Felipe Adrian Moreno Vera

Explorando a Percepção de Segurança Urbana com Modelos de Visão-Linguagem e Contrafactuais

Rio de Janeiro

June, 2026

Felipe Adrian Moreno Vera

Explorando a Percepção de Segurança Urbana com Modelos de Visão-Linguagem e Contrafactuais

Tese submetida à Escola de Matemática Aplicada (FGV/EMAp) como requisito para a obtenção do grau de Doutor em Matemática Aplicada e Ciência de Dados.

Área de Pesquisa: Computação Urbana, IA Generativa, Modelos de Visão-Linguagem (VLMs) e Análise Visual.

Fundação Getúlio Vargas
Escola de Matemática Aplicada

Orientador PhD. Jorge Poco

Rio de Janeiro
June, 2026

Agradecimentos

Esta pesquisa faz parte de um programa de Doutorado na Fundação Getulio Vargas (FGV). Em primeiro lugar, gostaria de agradecer ao *Inner Circle*, também conhecido como a "geração de ouro", bem como ao meu orientador por sua orientação, ao Laboratório de Ciência de Dados para Visualização (Visual-DS) pelas discussões e pelo apoio, e à minha família pelo suporte contínuo, mesmo à distância.

Além disso, este trabalho recebeu apoio da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Brasil; da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil; da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); e da Escola de Matemática Aplicada da Fundação Getulio Vargas (FGV/EMAp −− VisualDS).

Quaisquer opiniões, resultados, conclusões ou recomendações expressos neste material são de responsabilidade dos autores e não refletem, necessariamente, as posições da FAPESP, da FAPERJ, da CAPES ou da FGV.

"Para minha familia,
quem sempre esteve me apoiando a conseguir minhas metas
:)"

Felipe A. Moreno

Declaração de Autoria

Declaro, por meio desta, que a tese submetida é de minha autoria e constitui um trabalho original. Todas as fontes utilizadas, diretas ou indiretas, estão devidamente citadas nas referências. Declaro, ainda, que esta tese não foi submetida a nenhuma outra instituição para a obtenção de qualquer título acadêmico.

—

Acrônimos

AI *Artificial Intelligence*

SLR *Systematic Literature Review*

SVI *Street View Imagery*

GSV *Google Street View*

PRISMA *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*

SVM *Support Vector Machine*

RankSVM *Ranking Support Vector Machine*

SVR *Support Vector Regressor*

RL *Reinforcement Learning*

GMM *Gaussian Mixture Models*

KNN *K-Nearest Neighbors*

LASSO *Least Absolute Shrinkage and Selection Operator*

LLMs *Large Language Models*

VLMs *Vision-Language Models*

Abstract

Over several decades, urban perception has emerged as an important research domain spanning urban planning, mental well-being, and perception mapping. Foundational theories such as Broken Windows have highlighted the relationship between the built environment and human behavior. More recently, researchers have leveraged street-view imagery together with auxiliary datasets, including crime statistics, health indicators, and demographic information, to model how people perceive urban spaces. Despite their success, existing approaches largely rely on similar machine learning pipelines and provide limited insight into the visual factors that influence perceived safety. Although these models can estimate perceptual attributes, their opaque decision-making processes hinder transparency and reduce their usefulness for informing urban design. Counterfactual image generation offers a promising direction for interpretability, but current methods often produce unrealistic or poorly controlled edits, limiting their practical value.

This thesis makes four contributions. First, it reviews the urban perception literature to identify methodological limitations and research opportunities that motivate the subsequent work. Second, it investigates which visual and semantic elements of street-view images are most consistently associated with perceived safety through **UrbanFormer**, a model that jointly predicts perception scores and explains its own decisions. Third, building on this analysis, it introduces **UrbanPD4k**, a dataset of manually annotated physical disorder elements, and **UrbanCFX**, a pipeline that translates counterfactual scene differences into human-readable recommendations, making model outputs more accessible to non-specialists. Fourth, it introduces **UrbanCounterViz**, a visual analytics system for interpretable and steerable exploration of urban perception that integrates generative modeling and perception analysis. Rather than relying on unconstrained image generation, the system grounds edits in real observed transformations and presents them as photorealistic, inspectable visualizations through interactive, human-in-the-loop exploration. Evaluations based on usage scenarios, expert interviews, and an exploratory user study provide evidence that UrbanCounterViz can produce edits perceived as grounded, interpretable, and credible, supporting its use as an early-stage reasoning tool for data-driven urban design and planning.

Keywords: Systematic review; PRISMA; Urban perception; Visual analytics; Guided image editing; Counterfactual images; Street-view imagery; Semantic segmentation.

Resumo

A percepção urbana tornou-se um importante campo de pesquisa que conecta planejamento urbano, bem-estar e análise da relação entre ambiente construído e comportamento humano. Estudos recentes utilizam imagens de visão de rua combinadas com dados auxiliares, como estatísticas criminais, indicadores de saúde e informações demográficas, para modelar como as pessoas percebem os espaços urbanos. No entanto, embora esses modelos estimem atributos perceptuais, seus processos de decisão permanecem pouco transparentes, limitando sua aplicação no planejamento urbano. Métodos de geração de imagens contrafactuais oferecem potencial para melhorar a interpretabilidade, mas ainda enfrentam desafios relacionados ao realismo e ao controle das alterações geradas.

Esta tese apresenta quatro contribuições: uma revisão da literatura sobre percepção urbana e suas limitações metodológicas; o **UrbanFormer**, uma análise dos elementos visuais e semânticos associados à percepção de segurança por meio de um modelo interpretável; o **UrbanPD4k**, um conjunto de dados anotado de elementos de desordem urbana, e o **UrbanCFX**, um pipeline para transformar diferenças entre cenas em recomendações compreensíveis; e o UrbanCounterViz, um sistema de análise visual que integra modelagem generativa e análise perceptual. O sistema identifica diferenças visuais associadas a mudanças na percepção urbana e apresenta transformações realistas e interpretáveis baseadas em cenas reais observadas. Avaliações com especialistas e usuários indicam que o UrbanCounterViz pode apoiar processos de raciocínio e tomada de decisão no projeto e planejamento urbano orientados por dados.

Palavras-chave: Revisão sistemática; PRISMA; Percepção urbana; Análise visual; Edição guiada de imagens; Imagens contrafactuais; Imagens de ruas; Segmentação semântica; Percepção humana de segurança.

List of Figures

List of Tables

Lista de Prompts

Contents

1 Introdução

"*Cities are designed to shape and influence the lives of their inhabitants*" (Lindal and Hartig, 2013). Vários estudos têm demonstrado que a aparência visual das cidades desempenha um papel central na percepção e na reação humana ao ambiente. Um exemplo notável é a "*Broken Window theory*" (Wilson and Kelling, 1982), que sugere que sinais visuais de desordem ambiental, como janelas quebradas, carros abandonados, lixo e pichações, podem induzir resultados sociais negativos e aumentar os níveis de criminalidade. Essa teoria influenciou significativamente as estratégias de políticas públicas, levando a táticas policiais agressivas para controlar as manifestações de desordem social e física. Por exemplo, no estudo realizado na cidade de Nova York (Keizer et al., 2008; Hinkle and Weisburd, 2008), foram conduzidos experimentos sociais sobre a qualidade de vida percebida nas ruas. Esses experimentos compararam lugares "*impeccable*", como shopping centers (paredes limpas, organizados e silenciosos), com outros lugares onde havia a presença de pichações, ruas antigas ou negligenciadas e lixo nas ruas, concluindo que, em lugares onde "*rules are violated*", a longo prazo, as normas sociais não são respeitadas.

De forma semelhante, outros estudos demonstraram que a aparência visual dos espaços urbanos afeta o estado psicológico de seus habitantes (Lindal and Hartig, 2013). Por exemplo, uma avaliação psicológica demonstrou que a presença de áreas verdes nas cidades tende a produzir sensações positivas em seus habitantes, como segurança, relaxamento, tranquilidade etc. (Ulrich, 1979). Por outro lado, por meio do estudo de 40 relatórios psicológicos baseados em pesquisas e estudos sobre os estados mentais de seus habitantes, também se deduziu que a desordem urbana induz sofrimento psicológico, estresse e medo constante (Sampson et al., 2002). Além disso, pichações e edifícios em condições precárias ou de abandono também se mostraram diretamente relacionados à percepção de insegurança, determinando que uma aparência visual negligenciada influencia negativamente o ambiente (por exemplo, pichações, lixo espalhado pelas ruas, falta de limpeza no ambiente etc.) (Schroeder and Anderson, 1984).

Portanto, por meio de diversos estudos sobre o impacto da aparência visual de uma cidade em seus habitantes, torna-se vital compreender as percepções e avaliações dos espaços urbanos pelas pessoas. Nesse sentido, vários estudos foram realizados sobre como a cidade e sua aparência visual influenciam o comportamento na cidade. Por exemplo, em "*The Image of the City*" (Lynch, 1984), cidades (como Boston, Jersey e Los Angeles) foram divididas em regiões de importância (com base em dados sobre criminalidade, sociedade, seções urbanas ou não urbanas etc.), gerando mapas mentais sobre as características comuns dessas cidades. Em "*The Death and Life of Great American Cities*" (Jacobs, 1961), indica-se que os elementos de cada cidade se distinguem entre centenas, milhares

ou milhões de outros artefatos devido a suas formas, tamanhos, cores etc. únicos.

Com base em estudos anteriores, iniciou-se uma tendência no aspecto psicológico de estudar e avaliar a percepção dos habitantes em relação aos elementos visuais da cidade. No trabalho realizado por (Nasar, 1998), fortemente relacionado a encontrar aquelas regiões ou áreas que eram mais agradáveis aos cidadãos, demonstrou-se que, na maioria das avaliações, predominavam áreas verdes, ruas temáticas, espaços abertos, shopping centers e aeroportos. Além disso, as áreas avaliadas como "*not pleasant*" pelos habitantes eram edifícios com estilos pouco atraentes, pichações, parques sem forma estabelecida e lugares abandonados. Adicionalmente, outros estudos relacionados à desordem da cidade (Skogan, 1992) focaram na presença de lixo nas ruas, edifícios abandonados e carros estacionados em esquinas desertas, que contribuem para a percepção de falta de controle, medo e insegurança na cidade.

Outros estudos buscaram compreender a correlação entre o comportamento criminal e a aparência visual da cidade (Skogan, 1992). Esses estudos exploram a sensação de insegurança vinculada à prevalência de crimes e delinquência em espaços públicos e ainda se baseiam em avaliações presenciais ou observações de campo para coletar dados sobre características do ambiente construído (Rzotkiewicz et al., 2018; Wilson et al., 2012). Esses estudos sobre taxas de criminalidade em diferentes cidades levaram ao desenvolvimento de diversas aplicações, incluindo mapas de criminalidade, estatísticas de dados e aplicações que preveem tendências criminais em áreas específicas (Stalidis et al., 2018), além de estudos sobre o impacto da aparência visual e da taxa de criminalidade na percepção urbana das cidades (Andersson et al., 2017).

Nos últimos anos, com o avanço de ferramentas e *frameworks* para analisar diferentes tipos de informação (por exemplo, imagens, numéricas, temporais) e a evolução de técnicas como o aprendizado profundo (*deep learning*), uma quantidade notável de bancos de dados complexos foi criada com o propósito de prever a percepção urbana. Alguns trabalhos coletam dados por meio de websites e pesquisas online com usuários, como o MIT Media Lab Place Pulse "*Which looks more safety?*" (Salesses et al., 2013), *scenec-or-mot* (Seresinhe et al., 2017), "*What makes London beauty?*" (Quercia et al., 2014), City-SAFE (Costa et al., 2019), entre outros. Estudos semelhantes quantificam a vegetação e identificam áreas verdes e sua influência na percepção urbana (Li et al., 2015b). Além disso, alguns estudos analisam a relação entre níveis de violência, a presença de árvores e o índice de desenvolvimento humano (Porzi et al., 2015; Arietta et al., 2014; Li et al., 2015a), a presença de pichações e sua correlação com a percepção de segurança urbana (Tokuda et al., 2019; Lavi. et al., 2022), e dividem as cidades de acordo com os tipos de objetos mais frequentes (por exemplo, árvores, lixo, edifícios, cercas, pichações etc.) e a respectiva percepção de segurança (Zhang et al., 2018b; Min et al., 2020).

Ao longo desta dissertação, o termo *segurança* refere-se à *segurança percebida*: um

julgamento sobre o quão segura uma cena de imagem de rua aparenta ser para observadores humanos com base em suas características visuais. Ele não deve ser interpretado como uma medida objetiva de criminalidade, risco físico, segurança no trânsito ou segurança pública real. Essa distinção é central para o enquadramento metodológico e ético do trabalho, uma vez que os escores utilizados ao longo da dissertação são derivados de comparações pareadas obtidas por *crowdsourcing*, e não de registros oficiais de crimes ou incidentes.

No entanto, prever a segurança percebida representa apenas um aspecto do desafio. Para uma análise e um projeto urbano eficazes, também é necessário compreender como uma cena poderia ser plausivelmente modificada para alterar a percepção. As abordagens computacionais existentes normalmente se baseiam em explicações baseadas em atributos ou em resumos semânticos (Pan et al., 2024; Ma et al., 2025; Wu et al., 2025; Min et al., 2020; Song et al., 2025). Embora esses métodos esclareçam o comportamento do modelo, eles oferecem suporte limitado ao raciocínio sobre modificações de cena orientadas ao planejamento e visualmente plausíveis. Além disso, as abordagens de edição de imagens contrafactuais e generativas, cada vez mais impulsionadas por *Vision-Language Models* (VLMs), oferecem uma direção promissora ao ilustrar como uma imagem poderia ser transformada para alterar o resultado de um modelo (Gomez et al., 2020; Augustin et al., 2022; Jeanneret et al., 2022; Hu et al., 2023, 2025; Tan et al., 2025). Contudo, esses métodos frequentemente otimizam as edições com base nas respostas do modelo ou em geração aberta baseada em *prompts*, em vez de fundamentá-las em transformações observadas em cenas urbanas reais. Consequentemente, podem produzir resultados visualmente impressionantes, mas difíceis de interpretar como intervenções urbanas realistas.

Essa limitação evidencia a necessidade de abordagens que conectem a modelagem preditiva a transformações urbanas interpretáveis e contextualmente fundamentadas. Em particular, os analistas precisam de ferramentas que apoiem o raciocínio sobre como mudanças na composição semântica, na organização espacial e na variação influenciam a segurança percebida, permanecendo, ao mesmo tempo, consistentes com condições urbanas realistas. A visualização e a análise visual (*visual analytics*) são bem adequadas a esse desafio, pois permitem a exploração integrada do conteúdo da imagem, da estrutura semântica, do comportamento do modelo e do conhecimento de domínio. Pesquisas anteriores demonstraram o valor de sistemas interativos para a análise de imagens urbanas e a interpretação de aprendizado de máquina por meio de comparação, visões vinculadas e análise com humano no processo (*human-in-the-loop*) (Shen et al., 2017; Sacha et al., 2019; Collaris and van Wijk, 2020; La Rosa et al., 2023; Choo and Liu, 2018). Entretanto, essas abordagens apoiam principalmente a interpretação de dados e saídas de modelos existentes, oferecendo suporte limitado à exploração sistemática de transformações urbanas hipotéticas, porém semanticamente plausíveis, e de seus efeitos previstos sobre a percepção.

Neste trabalho, apresentamos o UrbanCounterViz (detalhado nos Capítulos 6 e 7), um sistema de análise visual para a exploração sistemática e transparente da

percepção de segurança urbana por meio de VLMs e contrafactuais de imagem guiados semanticamente. Em vez de apenas detectar elementos urbanos ou depender de geração aberta de imagens, nossa abordagem constrói *edições urbanas contrafactuais fundamentadas* a partir de *diferenças semanticamente significativas* identificadas entre imagens reais com diferentes escores de percepção. Nesse sentido, nosso objetivo não é apenas produzir exemplos contrafactuais que alterem a saída de um modelo, mas derivar recomendações que sejam sustentadas por mudanças reais e, portanto, mais bem fundamentadas como evidência analítica.

Antes de perguntar como a segurança urbana percebida poderia ser explicada ou tratada, é necessário primeiro compreender o que já se sabe sobre ela — que é exatamente o propósito da revisão sistemática que abre esta dissertação (Moreno-Vera and Poco, 2026). Essa investigação revela um padrão consistente: a pesquisa tornou-se notavelmente boa em estimar o quão segura uma rua aparenta ser, mas comparativamente pouco dela explica por quê, ou oferece uma forma de agir sobre essa explicação. É nessa lacuna que o presente trabalho começa.

Compreender a segurança urbana percebida, sob essa ótica, não é uma única questão, mas uma sequência de questões cada vez mais concretas. A primeira é simplesmente **quais elementos visuais de uma rua** — sua vegetação, seus muros, a desordem ou a conservação visível nela — **estão associados ao quão segura ela é julgada**. Responder a isso de uma forma em que se possa confiar exige um modelo que preveja e explique a si mesmo ao mesmo tempo (Moreno-Vera et al., 2024). Mas saber que um elemento importa diz pouco sobre como esse conhecimento poderia informar mudanças plausíveis na cena: **se substituir um muro nu por um bem conservado, ou remover um cabo aéreo, mudaria de forma plausível como um lugar é sentido**. Transformar essa compreensão em uma recomendação concreta e legível por humanos é a tarefa estudada em Moreno-Vera et al. (2025). E mesmo uma recomendação bem formulada permanece abstrata até que possa ser vista — um analista que raciocina sobre uma rua precisa, em última instância, olhar para ela, e não apenas ler sobre como ela poderia mudar. Fechar essa última distância, de uma mudança descrita a uma mudança visível e inspecionável fundamentada em cenas urbanas reais, é o que o UrbanCounterViz, apresentado nos Capítulos 6 e 7, se propõe a fazer.

1.1 Declaração da tese

Compreender a segurança urbana percebida exige conectar três coisas que a pesquisa tratou, em grande parte, de forma separada: **quais elementos visuais** de uma cena moldam o quão segura ela parece, **quais mudanças específicas** nesses elementos significariam de forma plausível, e **se tais mudanças podem ser mostradas** e inspecionadas em vez de meramente afirmadas. Esta dissertação segue essa conexão do início ao fim, e

cada capítulo que se segue retoma a questão que o anterior deixou em aberto.

Em outras palavras, esta tese argumenta que a segurança urbana percebida não deve ser tratada apenas como um problema de predição, mas como um problema de análise visual contrafactual fundamentada: compreendê-la exige mais do que modelos preditivos precisos — exige ferramentas nas quais os modelos identifiquem elementos visuais perceptualmente relevantes, os traduzam em mudanças plausíveis de cena e permitam que os analistas inspecionem essas mudanças como transformações visuais baseadas em evidências.

1.2 Objetivos

Os objetivos deste estudo são os seguintes:

Objetivo geral

Este trabalho investiga a segurança urbana percebida a partir de *Street View Imagery* (SVI) por meio de um *pipeline* orientado a dados que integra modelagem de percepção visual, representação semântica, raciocínio contrafactual, VLMs e análise visual para conectar predição, explicação e intervenção visual fundamentada.

Objetivos específicos

Os objetivos específicos deste trabalho seguem a mesma progressão da declaração da tese apresentada acima, movendo-se do mapeamento do que já é conhecido, à identificação do que importa, à explicação do que poderia mudar, até tornar essa mudança visível.

1. **Mapear o estado da pesquisa em percepção urbana.** Conduzir uma revisão sistemática da análise de percepção urbana utilizando SVI, identificando as principais abordagens, conjuntos de dados, técnicas computacionais e lacunas metodológicas na literatura, e usar essas lacunas para motivar os objetivos restantes.

2. **Identificar os elementos visuais que moldam a segurança percebida.** Investigar quais elementos visuais e semânticos específicos de uma SVI estão mais consistentemente associados à segurança percebida, utilizando um modelo que prevê e explica suas próprias decisões de forma conjunta.

3. **Traduzir essas associações em recomendações legíveis por humanos.** Determinar quais mudanças específicas nos elementos visuais de uma cena mudariam de forma plausível sua segurança percebida, e expressar essas mudanças como recomendações em linguagem natural, em vez de saída numérica ou vetorial.

4. **Tornar as mudanças plausíveis visíveis e inspecionáveis.** Projetar e desenvolver um sistema de análise visual que fundamente as mudanças de cena propostas em imagens urbanas reais e observadas, as renderize como edições fotorrealistas e permita que um analista inspecione, rastreie e interrogue todo o processo de transformação.

1.3 Contribuições

Esta dissertação apresenta quatro contribuições, cada uma correspondendo a um dos objetivos acima.

Contribuição 1. Uma revisão sistemática da literatura sobre pesquisa computacional em percepção urbana utilizando SVI, identificando quatro lacunas metodológicas persistentes — percepção limitada sobre os fatores visuais, explicações estáticas e indiretas, edições generativas irrealistas e a ausência de ferramentas interativas com humano no processo (*human-in-the-loop*) — que motivam as contribuições restantes desta tese. Este trabalho foi submetido ao periódico *IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI)* como: **Moreno-Vera, Felipe**, e Jorge Poco. "*From Vision-centric to Multimodal: A Systematic Review of Street View–Based Urban Perception*" (Moreno-Vera and Poco, 2026).

Contribuição 2. UrbanFormer (Capítulo 4), um modelo de classificação que funde atributos visuais de CNN com um vetor explícito de composição semântica, combinado a uma metodologia de explicação que interseca a atenção do Grad-CAM com máscaras de segmentação no espaço da imagem. Esta contribuição identifica quais elementos visuais moldam mais fortemente a segurança percebida e estabelece o vocabulário visual sobre o qual as contribuições restantes são construídas. Este trabalho foi publicado na *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT '24)* como: **Moreno-Vera, Felipe**, Bruno Brandoli e Jorge Poco. "*What Makes a Place Feel Safe? Analyzing SVIs to Identify Relevant Visual Elements*" (Moreno-Vera et al., 2024).

Contribuição 3. UrbanPD4k e **UrbanCFX** (Capítulo 5): um conjunto de dados de 3.659 SVIs do Rio de Janeiro com anotações em nível de pixel de 13 elementos de desordem física, e um *pipeline* que gera vetores de diferença contrafactuais e os traduz em linguagem natural legível por humanos utilizando um *Vision-Language Models* (VLMs). Esta contribuição transforma um sinal numérico, voltado ao modelo, em uma recomendação em texto sobre a qual um não especialista pode agir diretamente. Este trabalho foi publicado na *IEEE International Conference on Big Data (BigData '25)* como: **Moreno-Vera, Felipe**, Andres De-la-Puente e Jorge Poco. "*UrbanPhysicalDisorder-4K: Understanding Urban Perception via Counterfactuals and Street View Signs of Physical Disorder*" (Moreno-Vera et al., 2025).

Contribuição 4. UrbanCounterViz (Capítulos 6–7), um sistema de análise

visual que fundamenta edições contrafactuais de cena em diferenças semanticamente significativas entre cenas urbanas reais e visualmente semelhantes, e as renderiza como transformações fotorrealistas e inspecionáveis por meio de uma interface interativa, com humano no processo (*human-in-the-loop*). Junto ao próprio sistema, esta contribuição inclui um fluxo de trabalho reutilizável de análise visual interpretável que formaliza, independentemente de qualquer implementação específica, como as predições do modelo, as mudanças semânticas de cena e as transformações visuais fundamentadas podem ser sistematicamente conectadas.

Table 1 – Publicações produzidas durante o doutorado e sua função nesta dissertação.

Work	**Chapter**	**Status**	**Role**	**Function in the dissertation**
Systematic literature review	3	Submitted to IEEE TETCI	Core	Maps the field and identifies the gaps that motivate the thesis.
UrbanFormer (Moreno-Vera et al., 2024)	4	Published at WI-IAT 2024	Core	Identifies visual and semantic elements associated with perceived safety.
UrbanPD4k / UrbanCFX (Moreno-Vera et al., 2025)	5	Published at BigData 2025	Core	Adds disorder annotations and translates counterfactual vectors into language.
UrbanCounterViz	6–7	To be submitted to EuroVis 2027	Core	Makes grounded counterfactual changes visible, inspectable, and interactive.
VLM vs. human ratings (Moreno-Vera and Poco, 2025)	—	Published in IJCNN 2025	Complementary	Informs the VLM-based generation strategy but is not a core chapter.

Ao longo deste doutorado, também publicamos pesquisas adicionais que não fazem parte diretamente das contribuições desta dissertação. Em particular, o artigo **Moreno-Vera, Felipe**, e Jorge Poco. "*Assessing Urban Environments with VLMs: AI-Generated Ratings vs Human Evaluations*", publicado na *IEEE International Joint Conference on Neural Networks (IJCNN '25)* (Moreno-Vera and Poco, 2025), investiga o uso de VLMs para a avaliação de percepção urbana e complementa a agenda de pesquisa mais ampla desenvolvida durante o programa de doutorado.

A Tabela 1 distingue os trabalhos que constituem as contribuições centrais desta dissertação daqueles de pesquisa complementar produzidos durante o período do doutorado. As contribuições centrais definem a progressão da tese, enquanto o trabalho complementar informou escolhas metodológicas, mas não é desenvolvido como um capítulo independente.

1.4 Organização

Esta dissertação está organizada em três partes principais. A primeira, compreendendo os Capítulos 1 a 3, estabelece a base conceitual e empírica: motiva o problema, apresenta o embasamento técnico necessário para acompanhar o restante do trabalho e mapeia o estado atual da pesquisa em percepção urbana por meio de uma revisão

sistemática da literatura. A segunda parte, compreendendo os Capítulos 4 a 7, apresenta as três contribuições empíricas e de sistema desenvolvidas durante o doutorado — sendo a revisão sistemática do Capítulo 3 a quarta contribuição, de caráter metodológico — seguindo a mesma progressão introduzida neste capítulo: da identificação de quais elementos visuais moldam a segurança percebida, à descrição de mudanças plausíveis a eles em linguagem simples, até tornar essas mudanças visíveis e inspecionáveis por meio do UrbanCounterViz. A terceira parte, compreendendo os Capítulos 8 e 9, reúne essas contribuições, reflete sobre suas limitações e delineia direções para trabalhos futuros.

O Capítulo 2 apresenta o embasamento técnico — segmentação semântica, explicação contrafactual e VLMs — necessário para acompanhar as contribuições apresentadas mais adiante. O Capítulo 3 relata uma revisão sistemática da literatura que mapeia como a pesquisa em percepção urbana utilizando SVI evoluiu, e identifica as quatro lacunas metodológicas que motivam esta dissertação.

O Capítulo 4 retoma a primeira dessas lacunas, questionando quais elementos visuais estão mais consistentemente associados à segurança percebida. Este trabalho foi apresentado na *IEEE/WIC/ACM International Conference on Web Intelligence* (WI-IAT '24) (Moreno-Vera et al., 2024). O Capítulo 5 baseia-se nisso ao traduzir tais associações em recomendações concretas e legíveis por humanos, apresentando ao longo do caminho um conjunto de dados de anotações de desordem física para cenas de rua do Rio de Janeiro; este trabalho foi apresentado na *IEEE International Conference on Big Data* (BigData '25) (Moreno-Vera et al., 2025). Um estudo relacionado, apresentado na *IEEE International Joint Conference on Neural Networks* (IJCNN '25) (Moreno-Vera and Poco, 2025), comparou avaliações de ambientes urbanos geradas por IA e por humanos utilizando VLMs; embora não desenvolvido em um capítulo próprio, suas conclusões informaram a estratégia de geração baseada em VLM adotada no Capítulo 6.

Os Capítulos 6 e 7 apresentam o UrbanCounterViz, o sistema de análise visual que fecha a distância entre uma mudança descrita e uma mudança visível, fundamentando cada edição em cenas urbanas reais e observadas, em vez de geração aberta. O Capítulo 6 situa o UrbanCounterViz em relação aos sistemas de análise visual e de edição generativa mais próximos a ele antes de detalhar seu projeto e implementação, enquanto o Capítulo 7 o avalia por meio de cenários de uso, entrevistas com especialistas e um estudo com usuários mais amplo.

Por fim, o Capítulo 8 sintetiza os achados de todas as contribuições e discute as limitações e as direções futuras do trabalho como um todo, enquanto o Capítulo 9 conclui a dissertação retomando os objetivos definidos neste capítulo e refletindo sobre as implicações mais amplas da tese para a percepção urbana, a IA interpretável e a análise visual.

2 Fundamentação teórica

Este capítulo apresenta apenas os conceitos necessários para acompanhar as contribuições da dissertação: tarefas de aprendizado para SVI, interpretabilidade e explicações contrafactuais, VLMs multimodais e percepção urbana computacional. Em vez de fornecer uma revisão geral de aprendizado de máquina, ele se concentra nos componentes metodológicos que são posteriormente combinados no UrbanFormer (Capítulo 4), no UrbanCFX (Capítulo 5) e no UrbanCounterViz (Capítulos 6–7). O capítulo está organizado em quatro seções. A Seção 2.1 apresenta os paradigmas de aprendizado de máquina, as formulações de problemas e as tarefas de aprendizado relevantes para a análise de SVI. A Seção 2.2 aborda a interpretabilidade, os métodos de explicação baseados em gradiente e o raciocínio contrafactual — os fundamentos metodológicos para as contribuições de explicação dos Capítulos 4 e 5. A Seção 2.3 apresenta o aprendizado de representação multimodal e os VLMs (VLMs), o fundamento metodológico para a etapa de tradução em linguagem natural do Capítulo 5 e para o *pipeline* de edição de imagem em dois estágios do Capítulo 6. A Seção 2.4 apresenta os conceitos de percepção urbana, os protocolos de coleta de percepção e os métodos de quantificação utilizados ao longo desta tese.

2.1 Aprendizado de Máquina e Aprendizado Profundo

O aprendizado de máquina e o aprendizado profundo representam dois ramos poderosos da *Artificial Intelligence* (AI) que revolucionaram a forma como abordamos problemas complexos em diversos domínios. Em sua essência, tanto o aprendizado de máquina quanto o aprendizado profundo visam permitir que os computadores aprendam a partir de dados e façam predições ou decisões sem serem explicitamente programados para cada tarefa (Bengio, 2007).

O aprendizado de máquina engloba uma variedade de algoritmos e técnicas, desde algoritmos clássicos, como regressão linear e árvores de decisão, até abordagens mais avançadas, como máquinas de vetores de suporte e florestas aleatórias. O aprendizado de máquina permite que os computadores aprendam padrões e relações a partir dos dados e tomem decisões ou façam predições informadas. Além disso, o aprendizado profundo surgiu como uma abordagem particularmente poderosa para tarefas que envolvem dados complexos e de alta dimensionalidade, como imagens, áudio e texto. Os modelos de aprendizado profundo são construídos a partir de arquiteturas de redes neurais que aprendem representações hierárquicas diretamente dos dados. Essas redes neurais consistem em camadas interconectadas de neurônios que processam e transformam os dados, aprendendo representações hierárquicas de características diretamente de entradas brutas (Goodfellow

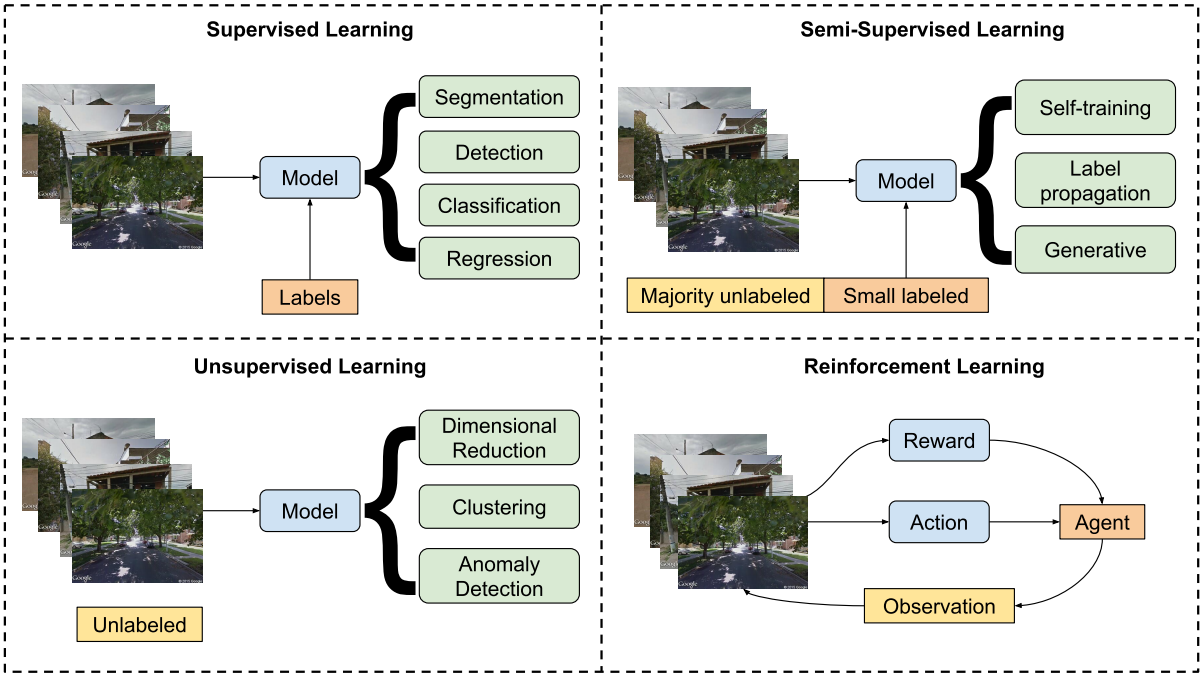

Figure 1 – Principais problemas de aprendizado aplicados a dados de SVI como entrada; apresentamos e especificamos exemplos de métodos para configurações de aprendizado supervisionado, não supervisionado, semissupervisionado e por reforço. Fonte: o autor.

et al., 2016), por meio de técnicas como CNNs para reconhecimento de imagens, RNNs para dados sequenciais e transformers para processamento de linguagem natural. Em ambos os casos, podemos classificar essas abordagens em quatro tipos de aprendizado: supervisionado, não supervisionado, semissupervisionado e por reforço. Além disso, cada tipo de aprendizado é categorizado em problemas de aprendizado e tarefas de aprendizado.

Tipo de aprendizado

O tipo de aprendizado refere-se à abordagem ou paradigma abrangente que caracteriza como um algoritmo de aprendizado de máquina aprende a partir dos dados. Os principais tipos de aprendizado em aprendizado de máquina são:

- **Aprendizado Supervisionado**: No aprendizado supervisionado, o sistema é treinado em um conjunto de dados rotulado, em que cada exemplo do conjunto de dados é associado a um rótulo ou alvo correspondente. O sistema aprende a mapear os dados de entrada para os rótulos de saída minimizando uma função de perda predefinida, otimizando assim sua acurácia preditiva. Exemplos de tarefas de aprendizado supervisionado incluem classificação (atribuir rótulos a entradas) e regressão (prever valores contínuos).

- **Aprendizado Não Supervisionado**: O aprendizado não supervisionado envolve treinar o sistema em um conjunto de dados não rotulado, em que os dados de

entrada carecem de rótulos ou alvos explícitos. Em vez disso, o sistema aprende a descobrir padrões, estruturas ou relações dentro dos dados, como agrupamentos ou representações latentes. Tarefas comuns de aprendizado não supervisionado incluem agrupamento (agrupar pontos de dados semelhantes) e redução de dimensionalidade (comprimir os dados preservando suas características essenciais).

- **Aprendizado Semissupervisionado**: O aprendizado semissupervisionado combina elementos dos aprendizados supervisionado e não supervisionado, em que o sistema é treinado em um conjunto de dados contendo uma pequena proporção de exemplos rotulados juntamente com uma quantidade maior de dados não rotulados. O sistema utiliza os dados rotulados para guiar seu processo de aprendizado, ao mesmo tempo em que explora a informação adicional presente nos dados não rotulados para aprimorar seu desempenho.

- **Aprendizado por Reforço**: O aprendizado por reforço lida com o treinamento de um agente para interagir com um ambiente a fim de maximizar recompensas cumulativas. O agente aprende por tentativa e erro, recebendo retorno na forma de recompensas ou penalidades com base em suas ações. Ao explorar diferentes estratégias e aprender com as consequências, o agente melhora gradualmente suas capacidades de tomada de decisão e aprende a atingir seus objetivos de forma eficiente.

A Figura 1 apresenta os quatro tipos de aprendizado de máquina e os principais problemas de aprendizado aplicados a SVI. Nesse contexto, o aprendizado supervisionado usa rótulos para realizar classificação e regressão. O aprendizado não supervisionado descobre padrões dentro de dados não rotulados. O aprendizado semissupervisionado usa uma combinação de dados rotulados e não rotulados para melhorar o desempenho do modelo. O aprendizado por reforço envolve aprender por tentativa e erro, interagindo com um ambiente para atingir um objetivo.

Problema de aprendizado

Um problema de aprendizado refere-se a um objetivo ou tarefa específicos que um algoritmo de aprendizado de máquina busca resolver. Ele define o que o algoritmo deve aprender a partir dos dados. Os problemas de aprendizado são normalmente categorizados de acordo com o tipo de aprendizado envolvido e a natureza da tarefa. Fornecemos uma visão geral concisa dos problemas de aprendizado mais comumente utilizados identificados neste trabalho:

- **Classificação**: É um problema de aprendizado de máquina em que os objetivos são inferir uma classe ou categoria a partir de um conjunto de observações ou amostras.

Os dados de entrada são $X \in \mathcal{R}^{n \times d}$, onde n é o número de amostras e d é o número de características (Goodfellow et al., 2016). A saída $y \in \{1, 2, \ldots, c\}^n$ é conhecida, onde c representa as categorias. O objetivo é aprender um modelo $f : \mathcal{R}^d \rightarrow \{1, 2, \ldots, c\}$. A classificação é um problema de Aprendizado Supervisionado, uma vez que a entrada e a saída são conhecidas. Alguns exemplos são segmentação semântica, classificação de imagens, reconhecimento de cenas, compreensão de cenas e detecção de objetos.

- **Regressão**: Semelhante ao problema de classificação, é um problema de aprendizado supervisionado em que, para cada amostra, precisamos estimar um valor numérico em vez de uma classe ou categoria (Goodfellow et al., 2016). O modelo é $f : \mathcal{R}^d \rightarrow \mathcal{R}$, onde a entrada e a saída são $X \in \mathcal{R}^{n \times d}$ e $y \in \mathcal{R}$. Alguns exemplos são o método ridge, o *Least Absolute Shrinkage and Selection Operator* (LASSO), o ElasticNet, o *Support Vector Regressor* (SVR) e o *Support Vector Machine* (SVM), entre outros.

- **Aprendizado de ranqueamento**: Também conhecido como *learning-to-rank*, é uma tarefa de aprendizado supervisionado em aprendizado de máquina que visa treinar modelos para ranquear um conjunto de itens com base em sua relevância ou utilidade para uma dada consulta ou usuário (Sutton and Barto, 2018). Formalmente, dado um conjunto de dados $D = (x_i, y_i)_{i=1}^N$, onde $x_i \in X$ representa as características do i-ésimo item a ser ranqueado e $y_i \in Y$ representa seu ranqueamento. O modelo é $f : X \rightarrow Y$, que mapeia as características de entrada para os ranqueamentos. Alguns exemplos são o *Ranking Support Vector Machine* (RankSVM), o LambdaRank, o RankNet, o Hinge e o RankBoost.

- **Agrupamento**: O problema de agrupamento visa identificar grupos (chamados de *clusters*) usando um conjunto de amostras e suas características. Diferentemente dos problemas anteriores, o agrupamento é um problema de aprendizado Não Supervisionado, uma vez que temos informação apenas sobre as entradas e desconhecemos as saídas (Abu-Mostafa et al., 2012). O objetivo é aprender um modelo $f : \mathcal{R}^d \rightarrow \{1, 2, \ldots, g\}$ que mapeia cada amostra $X \in \mathcal{R}^d$ para um agrupamento $y \in \{1, 2, \ldots, g\}$. Alguns exemplos são o Kmeans, o *K-Nearest Neighbors* (KNN) e o *Gaussian Mixture Models* (GMM), entre outros.

- **Generativo**: Em aprendizado de máquina, Generativo refere-se a uma classe de modelos ou algoritmos que visam aprender a distribuição de probabilidade subjacente de um conjunto de dados. Esses modelos são treinados para gerar novas amostras de dados que sejam semelhantes às observadas nos dados de treinamento (Goodfellow et al., 2016). Formalmente, um modelo generativo G aprende a distribuição de probabilidade conjunta $P(X, Y)$ dos dados de entrada X e dos rótulos ou dados de saída correspondentes Y, onde X e Y podem ser vetores, imagens, sequências etc. Dada a distribuição aprendida, um modelo generativo pode então gerar novas amostras amostrando a distribuição de probabilidade condicional aprendida $P(X|Y)$.

- **Agente de política**: É um problema de aprendizado de máquina em que um agente aprende a tomar decisões interagindo com um ambiente para atingir um determinado objetivo. Em *Reinforcement Learning* (RL), o agente aprende por tentativa e erro, recebendo retorno na forma de recompensas ou penalidades com base em suas ações (Sutton and Barto, 2018). Dados um conjunto de estados S, um conjunto de ações possíveis, a função de recompensa $R : S \times A \to \mathcal{R}$ e uma função de transição $T : S \times A \times S \to [0, 1]$, que fornece a probabilidade de obter um novo estado. O objetivo é aprender uma política de ação $\pi : S \to A$ que decide qual ação $a = \pi(s) \in A$ deve ser tomada a fim de maximizar a recompensa final após $k \in \mathcal{N}$ passos.

Tarefa de aprendizado

Uma tarefa de aprendizado refere-se a uma instância ou aplicação específica de um problema de aprendizado dentro de um contexto particular. Frequentemente, envolve definir os dados de entrada, a saída desejada e as métricas de avaliação para avaliar o desempenho do algoritmo de aprendizado. As tarefas de aprendizado podem variar amplamente dependendo do domínio e do problema específico que está sendo abordado. Alguns exemplos são classificação de imagens, detecção de objetos, segmentação de imagens, análise de sentimentos, detecção de anomalias, regressão multivariada, reconhecimento de cenas e classificação por comitê (*ensemble*), entre outros.

Em resumo, o tipo de aprendizado descreve a abordagem abrangente para o aprendizado, o problema de aprendizado define o objetivo ou a tarefa a ser resolvida, e a tarefa de aprendizado especifica uma instância ou aplicação particular de um problema de aprendizado em um contexto específico. Compreender essas distinções ajuda a enquadrar e abordar de forma eficaz os problemas de aprendizado de máquina em diversos domínios.

2.2 Interpretabilidade e Explicabilidade

Interpretabilidade e explicabilidade são conceitos relacionados, porém distintos, na literatura de aprendizado de máquina (Lipton, 2018). A **interpretabilidade** é uma propriedade intrínseca de um modelo, referindo-se ao grau em que seus mecanismos internos podem ser diretamente compreendidos por um observador humano. A **explicabilidade**, por outro lado, é extrínseca: diz respeito à geração de artefatos *post-hoc* que resumem o comportamento do modelo em termos compreensíveis para humanos, sem necessariamente revelar o processo computacional subjacente. De forma crucial, um modelo pode ser explicável sem ser interpretável — uma rede neural profunda pode produzir mapas de atribuição fiéis enquanto suas representações aprendidas permanecem opacas.

(a) (b) (c)

Figure 2 – Predição: Cachorro; explicações produzidas por (a) CAM, (b) GBP e (c) Guided GradCAM. Fonte: (Selvaraju et al., 2017).

2.2.1 Métodos de explicação

Os métodos de explicação fornecem percepções sobre o comportamento e os processos internos de tomada de decisão dos modelos de aprendizado de máquina. Esses métodos podem ser amplamente categorizados em duas famílias: métodos *white-box*, que operam sobre modelos inerentemente transparentes, como classificadores lineares, e que são, portanto, diretos de analisar; e métodos *black-box*, que têm como alvo arquiteturas complexas com grande número de parâmetros — como as redes neurais profundas — onde o raciocínio interno não é diretamente acessível (Molnar, 2022). Técnicas representativas nesta última categoria incluem a **Convolution Visualization** (Zeiler and Fergus, 2014), os gradientes *smooth* (Ancona et al., 2017), os *saliency maps* (Simonyan et al., 2013) e os *class activation maps* (CAM) (Zhou et al., 2016), entre outros.

Para formalizar o cenário, seja $x \in \mathbb{R}^d$ o vetor de características de uma imagem de entrada, $S : \mathbb{R}^d \to \mathbb{R}^C$ um modelo que mapeia entradas para escores sobre C classes, e $E : \mathbb{R}^d \to \mathbb{R}^d$ uma função de explicação que produz um mapa de atribuição sobre as dimensões da entrada. Sob esse arcabouço, uma explicação baseada em gradiente assume a forma $E_{grad}(x) = \frac{\partial S}{\partial x}$, onde o gradiente captura a sensibilidade de cada dimensão de entrada à predição do modelo $S(x)$ dentro de uma vizinhança local de x (Adebayo et al., 2018). Os seguintes métodos baseados em gradiente se apoiam nesse fundamento:

- **Gradient Input**: Calcula o produto elemento a elemento da entrada e seu gradiente, expresso como $x \odot \frac{\partial S}{\partial x}$. Essa formulação suprime a difusão visual e mitiga o problema de saturação do gradiente (Shrikumar et al., 2016).
- **Integrated Gradients**: Acumula gradientes ao longo de um caminho reto de uma baseline $\bar{x}$ — representando a ausência de características — até a entrada x, resultando em $(x - \bar{x}) \int_0^1 \frac{\partial S(\bar{x}+\alpha(x-\bar{x}))}{\partial x} \, d\alpha$. A integração sobre o fator de escala α reduz a saturação do gradiente e satisfaz um conjunto de propriedades axiomáticas desejáveis (Sundararajan et al., 2017).

- **CAM (Class Activation Map)**: Produz um mapa de localização discriminativo de classe da forma $M_{cam} = \sum_k w_k^c F^k$, onde w_k^c denota o peso associado à classe c para o k-ésimo mapa de características convolucional, e $F^k = \frac{1}{Z}\sum_i \sum_j A_{ij}^k$ é o resultado da aplicação do Global Average Pooling (GAP) ao mapa de ativação A^k. Aqui, (i, j) indexam posições espaciais e Z é o número total de unidades espaciais (Zhou et al., 2016).

- **GradCAM**: Estende o CAM substituindo os pesos de classificação aprendidos por escores de importância derivados do gradiente. Especificamente, o peso de importância do neurônio é definido como $\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial M_c}{\partial A_{ij}^k}$, referido como a *partial linearization* do gradiente via GAP. O mapa de localização final é então calculado como $M_{gradcam} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$ (Selvaraju et al., 2017).

- **GBP (Guided Backpropagation)**: Modifica a retropropagação padrão suprimindo os gradientes negativos que fluem pelas ativações ReLU. O sinal na camada t para a amostra i é dado por $B_i^t = (f_i^t > 0) \times (B_i^{t+1} > 0) \times B_i^{t+1}$, onde $f_i^{t+1} = \text{ReLU}(f_i^t)$ é a ativação direta e $B_i^{t+1} = \frac{\partial f^{out}}{\partial f_i^{t+1}}$ é o gradiente a montante (Springenberg et al., 2014).

- **Guided GradCAM**: Combina a precisão espacial do GBP com as propriedades discriminativas de classe do GradCAM por meio da multiplicação elemento a elemento dos dois mapas de atribuição (Selvaraju et al., 2017).

- **SmoothGrad (SG)**: Reduz o ruído e a difusão visual nos *saliency maps* calculando a média das explicações obtidas sobre múltiplas perturbações ruidosas da entrada. Para um método de explicação E, a estimativa suavizada é $E_{sg} = \frac{1}{N}\sum_{i=1}^{N} E(x + g_i)$, onde $g_i \sim \mathcal{N}(0, \sigma^2)$ é um ruído gaussiano (Smilkov et al., 2017).

A Figura 2 ilustra as saídas do CAM, do GradCAM e do GBP em uma imagem de exemplo, destacando as diferenças qualitativas nos mapas de atribuição produzidos por cada método. Essas técnicas estão entre as ferramentas mais amplamente utilizadas para interpretar modelos treinados em dados visuais.

2.2.2 Métodos de interpretabilidade e contrafactuais

Enquanto os métodos de explicação focam em compreender *por que* um modelo produz uma predição específica — frequentemente por meio da atribuição *post-hoc* de características de entrada — a **interpretabilidade** refere-se ao grau em que um humano consegue compreender os mecanismos internos de um modelo (Lipton, 2018; Doshi-Velez and Kim, 2017). Essa distinção é sutil, mas importante: uma rede neural profunda altamente acurada pode ser explicável por meio de *saliency maps* e, ainda assim, permanecer em grande parte não interpretável, uma vez que as representações subjacentes codificadas em seus pesos não são diretamente acessíveis ao raciocínio humano. A interpretabilidade,

portanto, diz respeito à *transparência* do próprio modelo, e não à legibilidade de predições individuais.

Diversas linhas de pesquisa examinaram como conceitos de alto nível são representados dentro das redes neurais. As **Explicações baseadas em conceitos**, como o TCAV (Testing with Concept Activation Vectors) (Kim et al., 2018), sondam camadas intermediárias de uma rede para determinar se conceitos definidos por humanos — como "*stripes*" ou "*wheels*" — são linearmente separáveis no espaço de representação aprendido. Essa abordagem permite que os pesquisadores quantifiquem a sensibilidade das predições de um modelo a um dado conceito sem exigir modificações no próprio modelo. De forma semelhante, a **Network Dissection** (Bau et al., 2017) alinha unidades ocultas individuais a conceitos semânticos medindo a sobreposição entre as ativações das unidades e as máscaras de segmentação de conceitos visuais conhecidos, revelando que muitas unidades em redes neurais convolucionais correspondem a partes de objetos ou texturas interpretáveis. Uma linha de trabalho complementar foca nos **Probing Classifiers** (Alain and Bengio, 2018), em que classificadores lineares leves são treinados sobre as ativações de camadas individuais para avaliar quais propriedades linguísticas ou semânticas são codificadas em cada profundidade da rede. Essa abordagem foi amplamente adotada na análise de arquiteturas de VLMs.

As **Explicações contrafactuais** oferecem uma perspectiva diferente sobre o comportamento do modelo ao abordar a questão: *qual é a mudança mínima em uma entrada x que alteraria a predição do modelo?* Formalmente, dado um modelo f e uma classe predita $c = f(x)$, busca-se um contrafactual x' tal que $f(x') \neq c$ e a distância $\|x - x'\|$ seja minimizada de acordo com alguma métrica (Wachter et al., 2017). Diferentemente dos métodos de atribuição baseados em gradiente, os contrafactuais não exigem acesso aos gradientes do modelo e são, portanto, aplicáveis em cenários *black-box*. No domínio de imagens, o **DiCE** (Diverse Counterfactual Explanations) (Mothilal et al., 2020) gera múltiplos contrafactuais diversos para evitar a dependência excessiva de uma única explicação, enquanto as **Contrastive Explanations** (Dhurandhar et al., 2018) decompõem as explicações em *pertinent positives* (características que devem estar presentes para justificar a predição) e *pertinent negatives* (características cuja ausência é necessária).

Uma formulação particularmente relevante para modelos visuais é a dos **contrafactuais semânticos**, em que as mudanças são feitas em um espaço semântico latente em vez de no espaço de pixels, resultando em modificações mais coerentes e interpretáveis por humanos (Goyal et al., 2019). Essa linha de trabalho foi aplicada a tarefas de percepção visual em que compreender o papel dos atributos dos objetos — como cor, textura ou configuração espacial — é importante para auditar as decisões do modelo.

Apoiando-se nessa ideia, esta dissertação adota a noção de um **contrafactual fundamentado**: uma mudança proposta em uma cena urbana cuja direção é derivada

de diferenças observadas entre SVIs reais, em vez de uma geração irrestrita baseada em *prompts* ou de uma otimização puramente sintética no espaço de características ou latente. O fundamento, portanto, refere-se à origem empírica da edição — a transformação é ancorada em padrões visuais e semânticos que de fato ocorrem nos dados. Essa formulação, central para o sistema apresentado no Capítulo 6, troca parte da flexibilidade da geração aberta por plausibilidade e rastreabilidade: toda mudança proposta pode ser rastreada até uma diferença observada entre cenas reais.

Juntos, os métodos de interpretabilidade e as explicações contrafactuais complementam as abordagens de atribuição baseadas em gradiente descritas na seção anterior, fornecendo um panorama mais completo do comportamento do modelo: enquanto os métodos de atribuição destacam *onde* um modelo olha, os métodos de interpretabilidade revelam *o que* ele aprendeu, e os métodos contrafactuais caracterizam *como* ele decidiria de forma diferente — com os contrafactuais fundamentados adicionalmente restringindo esse *como* a mudanças observadas em dados reais.

2.3 Aprendizado Multimodal e Vision-Language Models (VLMs)

Os métodos discutidos até aqui operam sobre uma única modalidade de entrada, tipicamente uma imagem, e produzem um rótulo de classe, um escore contínuo ou um mapa de atribuição definido sobre essa mesma modalidade. Muitas das contribuições desta tese, no entanto, exigem um raciocínio que abrange duas modalidades simultaneamente: uma cena urbana expressa como uma imagem e uma descrição dessa cena, ou de uma mudança nessa cena, expressa em linguagem natural. Esta seção apresenta os conceitos de aprendizado de representação multimodal e de VLMs (VLMs) que fundamentam a tradução em linguagem natural das mudanças contrafactuais no Capítulo 5 e o *pipeline* de edição de imagem vision-language do Capítulo 6.

2.3.1 Aprendizado de representação multimodal

O **aprendizado multimodal** refere-se a uma família de métodos que processam conjuntamente informação de mais de uma modalidade — como visão, linguagem e áudio — a fim de aprender representações que capturem as relações entre elas (Baltrušaitis et al., 2018). Formalmente, seja $x_I \in \mathcal{X}_I$ uma imagem e $x_T \in \mathcal{X}_T$ uma sequência de texto. Um modelo multimodal aprende um par de codificadores,

$$f_I : \mathcal{X}_I \to \mathbb{R}^d, \qquad f_T : \mathcal{X}_T \to \mathbb{R}^d, \tag{2.1}$$

que mapeiam as entradas de cada modalidade para um espaço de embedding compartilhado de dimensionalidade d, de modo que pares imagem-texto semanticamente relacionados sejam posicionados próximos sob uma função de similaridade $\mathrm{sim}(\cdot, \cdot)$, tipicamente a

similaridade de cosseno,

$$\mathrm{sim}(x_I, x_T) = \frac{f_I(x_I)^\top f_T(x_T)}{\|f_I(x_I)\| \, \|f_T(x_T)\|}. \tag{2.2}$$

Esse espaço de embedding compartilhado é o que permite a um modelo posterior comparar, recuperar ou condicionar a geração entre modalidades, e constitui o bloco de construção básico dos VLMs descritos a seguir.

2.3.2 Vision-Language Models (VLMs)

Os **VLMs** estendem a formulação multimodal acima treinando o codificador de imagem f_I e o codificador de texto f_T conjuntamente em grandes corpora de dados imagem-texto pareados, mais comumente usando um objetivo contrastivo (Radford et al., 2021). Dado um lote de N pares imagem-texto $\{(x_I^i, x_T^i)\}_{i=1}^N$, o modelo é treinado de modo que os embeddings de pares correspondentes sejam aproximados, enquanto os embeddings de pares não correspondentes sejam afastados. Isso é obtido com a perda InfoNCE (Oord et al., 2018), aplicada simetricamente sobre imagens e texto,

$$\mathcal{L}_{I\to T} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(x_I^i, x_T^i)/\tau)}{\sum_{j=1}^{N} \exp\Big(\mathrm{sim}(x_I^i, x_T^j)/\tau\Big)}, \tag{2.3}$$

$$\mathcal{L}_{T\to I} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(x_T^i, x_I^i)/\tau)}{\sum_{j=1}^{N} \exp\Big(\mathrm{sim}(x_T^i, x_I^j)/\tau\Big)}, \tag{2.4}$$

onde τ é um parâmetro de temperatura aprendido que controla a acuidade da distribuição. O objetivo geral de treinamento é a média $\mathcal{L}_{\mathrm{VLM}} = \frac{1}{2}\left(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}\right)$. Modelos treinados sob essa formulação contrastiva, como o CLIP (Radford et al., 2021) e o ALIGN (Jia et al., 2021), aprendem representações que transferem bem para tarefas posteriores — incluindo classificação zero-shot e recuperação cross-modal — sem ajuste fino específico da tarefa, uma vez que uma nova classe ou consulta pode ser expressa simplesmente como um *prompt* de texto e comparada com os embeddings de imagem por meio da Equação (2.2).

Uma linha de trabalho complementar substitui o objetivo puramente contrastivo por uma arquitetura encoder-decoder ou transformer unificada, treinada em uma mistura de objetivos contrastivos, de legendagem e de correspondência, como o BLIP (Li et al., 2022a) e o BLIP-2 (Li et al., 2023). Esses modelos combinam um codificador de visão congelado ou levemente ajustado com um decodificador de linguagem, permitindo não apenas o alinhamento entre modalidades, mas também a geração de texto livre condicionada à entrada visual — uma capacidade que modelos puramente contrastivos, como o CLIP, não possuem, e sobre a qual se apoiam os VLMs generativos descritos a seguir.

2.3.3 VLMs generativos e que seguem instruções

Enquanto os modelos acima focam em *alinhar* as duas modalidades, uma segunda família de VLMs foca na *geração*: produzir texto livre condicionado à entrada visual, seguindo uma instrução em linguagem natural. Dada uma imagem x_I e uma instrução ou *prompt* x_T, esses modelos são tipicamente construídos sobre um modelo de linguagem autorregressivo que gera uma sequência de saída $y = (y_1, \ldots, y_L)$ token a token, condicionada a ambas as entradas,

$$p(y \mid x_I, x_T) = \prod_{t=1}^{L} p_\theta(y_t \mid y_{<t},\ f_I(x_I),\ x_T)\,, \tag{2.5}$$

onde a representação da imagem $f_I(x_I)$ — tipicamente uma sequência de tokens visuais ou embeddings de patches projetados no espaço de embedding do modelo de linguagem — é tratada como contexto adicional ao lado dos tokens gerados anteriormente $y_{<t}$. O treinamento procede minimizando a log-verossimilhança negativa de uma resposta de referência, seguindo a mesma formulação de entropia cruzada introduzida para o cenário de classificação na Seção 2.1, mas somada sobre a sequência:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{L} \log p_\theta(y_t \mid y_{<t},\ f_I(x_I),\ x_T)\,. \tag{2.6}$$

Modelos representativos treinados sob esse paradigma incluem o Flamingo (Alayrac et al., 2022), o LLaVA (Liu et al., 2023a) e sistemas proprietários de larga escala, como o GPT-4V (Achiam et al., 2023) e o Gemini (Team et al., 2023), que adicionalmente suportam interação multi-turno e de seguimento de instruções. No contexto desta tese, essa formulação generativa é o mecanismo pelo qual uma mudança numérica ou simbólica em uma cena urbana — por exemplo, um vetor de diferença contrafactual — é traduzida em uma recomendação em linguagem natural no Capítulo 5, e pelo qual uma instrução de edição é subsequentemente convertida em uma edição de imagem fotorrealista no *pipeline* de dois estágios do Capítulo 6, onde um VLMs se condiciona tanto à imagem de origem quanto a uma descrição estruturada da mudança pretendida para produzir a cena editada.

2.4 Percepção Urbana e IA

As imagens de *Street View* fornecem cobertura visual em larga escala de ambientes urbanos, e os métodos de aprendizado de máquina — particularmente as redes neurais profundas para detecção de objetos, segmentação semântica e compreensão de cenas — tornam possível analisar essas imagens em escala (Ibrahim et al., 2020; Zhang et al., 2024). Essa combinação fundamenta o campo da percepção urbana computacional, cujos protocolos de coleta de dados e de quantificação esta seção apresenta.

Em geral, os estudos de percepção urbana visam coletar e quantificar as percepções humanas sobre os ambientes urbanos. Muitos desses estudos se apoiam em plataformas

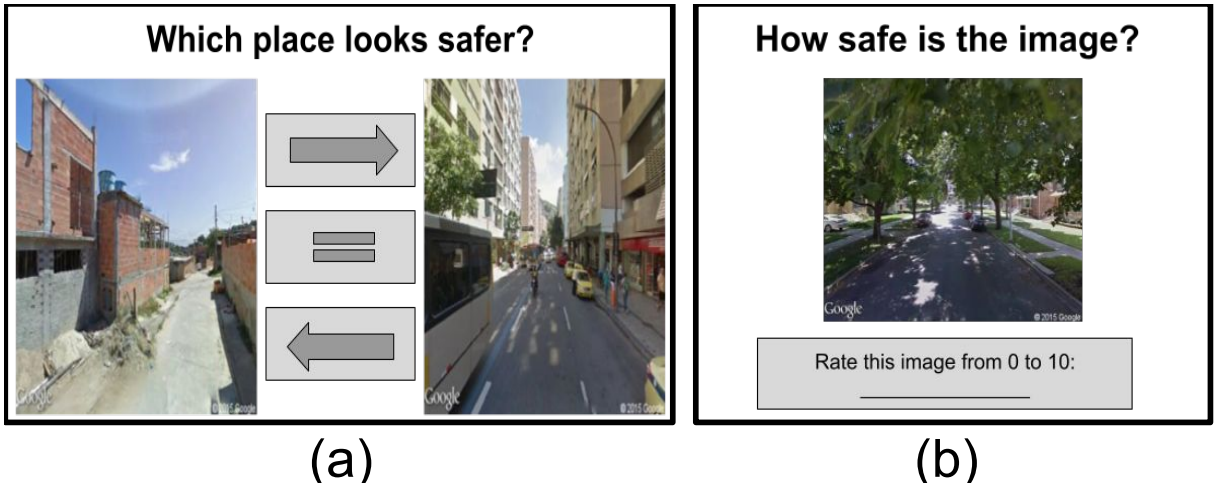

Figure 3 – Existem duas abordagens principais de coleta: (a) reunir informações de voluntários por meio de pesquisas, entrevistas ou questionários perguntando, por exemplo, *Which place looks safer?*, em que os voluntários escolhem entre duas imagens; ou (b) "*Rate the safety of this image*", solicitando que os voluntários avaliem uma imagem de 0 a 10. Fonte: o autor.

baseadas na web, como o Place Pulse (Salesses et al., 2013), o City-SAFE (Costa et al., 2019) e o Wmodi (Acosta and Camargo, 2018), para reunir dados de percepção em larga escala. Os participantes são tipicamente apresentados a pares de SVIs e solicitados a compará-los de acordo com atributos específicos, como segurança, beleza, vivacidade ou riqueza. Os conjuntos de dados resultantes geralmente incluem os pares de imagens, suas coordenadas geográficas e os julgamentos humanos correspondentes. Essas anotações podem então ser usadas para treinar modelos de aprendizado de máquina e de aprendizado profundo capazes de prever atributos perceptuais diretamente a partir de imagens urbanas, permitindo avaliações em larga escala dos ambientes urbanos sem exigir avaliação humana contínua.

Quatro termos recorrem ao longo desta dissertação e merecem definição explícita. A **segurança percebida** denota um julgamento sobre o quão segura uma cena de SVI aparenta ser para observadores humanos, derivado de comparações pareadas obtidas por *crowdsourcing*; ela é distinta da segurança objetiva, das taxas de criminalidade ou do risco físico (ver Capítulo 1). Uma **representação semântica** de uma cena é uma descrição estruturada de seu conteúdo em termos de categorias rotuladas — por exemplo, os vetores de razão de pixels por categoria produzidos pela segmentação semântica — em vez de valores de pixel brutos. A **análise visual** refere-se ao raciocínio analítico apoiado por interfaces visuais interativas, em que modelos computacionais e o julgamento humano são combinados por meio de visões coordenadas, e não de saídas estáticas. Por fim, um sistema é **human-in-the-loop** quando o analista participa ativamente do processo analítico — selecionando entradas, direcionando transformações e validando saídas — em vez de receber um resultado único e inegociável. A noção relacionada de *contrafactual fundamentado* é definida na Seção 2.2.

2.4.1 Coleta de percepção

A Figura 3 ilustra dois métodos principais para coletar julgamentos perceptuais: (a) escalas de avaliação e (b) comparações pareadas de ranqueamento.

Os **escores de avaliação**, comumente usando a escala Likert Likert (1932), pedem aos participantes que avaliem imagens em uma escala ordinal fixa (por exemplo, de 5 ou 10 pontos), com opções que variam de *strongly agree* a *strongly disagree*. Essa abordagem captura graus variados de percepção ou preferência, permitindo respostas mais matizadas do que simples escolhas binárias. Como resultado, os pesquisadores podem quantificar melhor as impressões subjetivas e comparar os resultados entre os participantes. O escore final da imagem é tipicamente calculado como a média de todas as respostas, conhecido como **Escore Médio**, que fornece um resumo direto da avaliação geral.

Os **escores de ranqueamento**, particularmente as comparações pareadas Thurstone (2017), apresentam duas imagens e pedem aos usuários que selecionem a preferida com base em um dado critério (por exemplo, segurança, beleza ou atratividade). Essa abordagem permite uma compreensão mais granular das preferências, uma vez que os indivíduos fazem comparações diretas em vez de avaliar as imagens isoladamente. A partir dessas comparações, um **escore de ranqueamento** final para cada imagem pode ser estimado usando métodos como o algoritmo "*strength of schedule*" Park and Newman (2005); Naik et al. (2014); Salesses et al. (2013), que leva em conta as taxas de vitória–derrota e a dificuldade do adversário; o sistema de pontuação Elo Ebtekar and Liu (2021), que atualiza os escores iterativamente com base nos resultados pareados; e o TrueSkill Minka et al. (2018), um arcabouço bayesiano que modela tanto as estimativas de habilidade quanto a incerteza.

2.4.2 Cálculo da percepção

Embora as avaliações forneçam escores diretamente, a comparação pareada é uma estratégia comum para quantificar percepções subjetivas a partir de dados visuais. Nessa abordagem, os anotadores selecionam qual imagem representa melhor um dado atributo (por exemplo, mais segura, mais bela ou mais viva). Esses resultados binários podem então ser usados por algoritmos de ranqueamento para estimar escores perceptuais latentes para cada imagem.

Sejam i e j duas imagens comparadas em uma tarefa pareada. O resultado da comparação é representado como:

$$S_i = \begin{cases} 1, & \text{se a imagem } i \text{ é preferida à imagem } j \\ 0, & \text{caso contrário} \end{cases} \tag{2.7}$$

De forma semelhante, o resultado para a imagem j é definido como:

$$S_j = 1 - S_i \tag{2.8}$$

Sistema de pontuação Elo

O sistema de pontuação Elo modela a probabilidade de que uma imagem seja preferida a outra como função da diferença entre suas pontuações. A cada imagem é atribuída uma pontuação escalar R, que é atualizada iterativamente após cada comparação. A probabilidade esperada de que a imagem i seja preferida à imagem j e a probabilidade esperada para a imagem j são definidas como:

$$E_i = \frac{1}{1 + 10^{\frac{R_j - R_i}{400}}} \tag{2.9a}$$

$$E_j = \frac{1}{1 + 10^{\frac{R_i - R_j}{400}}} \tag{2.9b}$$

Após observar o resultado da comparação, as pontuações são atualizadas de acordo com:

$$R'_i = R_i + K\left(S_i - E_i\right) \tag{2.10a}$$

$$R'_j = R_j + K\left(S_j - E_j\right) \tag{2.10b}$$

onde K é uma constante que controla a taxa de aprendizado das atualizações. Valores típicos de K variam de 16 a 32. O método Elo produz uma única pontuação determinística para cada imagem e é computacionalmente eficiente para conjuntos de dados pareados de larga escala.

Sistema de pontuação TrueSkill

O arcabouço TrueSkill (Minka et al., 2018) estende o sistema Elo modelando a incerteza na pontuação estimada de cada imagem. Em vez de um único valor de pontuação, cada imagem é representada por uma distribuição gaussiana:

$$s_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \tag{2.11}$$

onde μ_i representa a habilidade estimada (ou escore de percepção) e σ_i representa a incerteza associada a essa estimativa. O TrueSkill assume que o desempenho observado de uma imagem em uma comparação está sujeito a ruído:

$$p_i \sim \mathcal{N}(\mu_i, \sigma_i^2 + \beta^2) \tag{2.12}$$

onde β é um parâmetro que representa a variabilidade de desempenho. Para uma comparação pareada em que a imagem i é preferida à imagem j, a variância combinada é calculada como:

$$c^2 = 2\beta^2 + \sigma_i^2 + \sigma_j^2 \tag{2.13}$$

A diferença de desempenho padronizada é então definida como:

$$t = \frac{\mu_i - \mu_j}{c} \tag{2.14}$$

Duas funções auxiliares são usadas nas equações de atualização:

$$v(t) = \frac{\phi(t)}{\Phi(t)} \tag{2.15a}$$

$$w(t) = v(t)\left[v(t) + t\right] \tag{2.15b}$$

onde $\phi(t)$ e $\Phi(t)$ denotam a função densidade de probabilidade (PDF) e a função de distribuição acumulada (CDF) da distribuição normal padrão. Os valores médios atualizados são calculados como:

$$\mu_i' = \mu_i + \frac{\sigma_i^2}{c} v(t) \tag{2.16a}$$

$$\mu_j' = \mu_j - \frac{\sigma_j^2}{c} v(t) \tag{2.16b}$$

As variâncias atualizadas são:

$$\sigma_i'^2 = \sigma_i^2 \left[1 - \frac{\sigma_i^2}{c^2} w(t)\right] \tag{2.17a}$$

$$\sigma_j'^2 = \sigma_j^2 \left[1 - \frac{\sigma_j^2}{c^2} w(t)\right] \tag{2.17b}$$

Na prática, as imagens são ranqueadas usando uma estimativa conservadora de sua pontuação:

$$Q_i = \mu_i - 3\sigma_i \tag{2.18}$$

Essa formulação penaliza a alta incerteza e produz ranqueamentos estáveis, particularmente em conjuntos de dados de percepção obtidos por *crowdsourcing* com números variados de comparações por imagem.

Strength of Schedule

O Strength of Schedule (SOS) (Park and Newman, 2005) é um método usado para avaliar a dificuldade dos adversários de uma equipe, ajudando a fornecer uma avaliação mais justa do desempenho. Ele tipicamente considera os percentuais de vitória dos adversários e, em alguns casos, os adversários desses adversários. Essa métrica é amplamente usada em rankings esportivos para comparar equipes com diferentes níveis de competição. Assim,

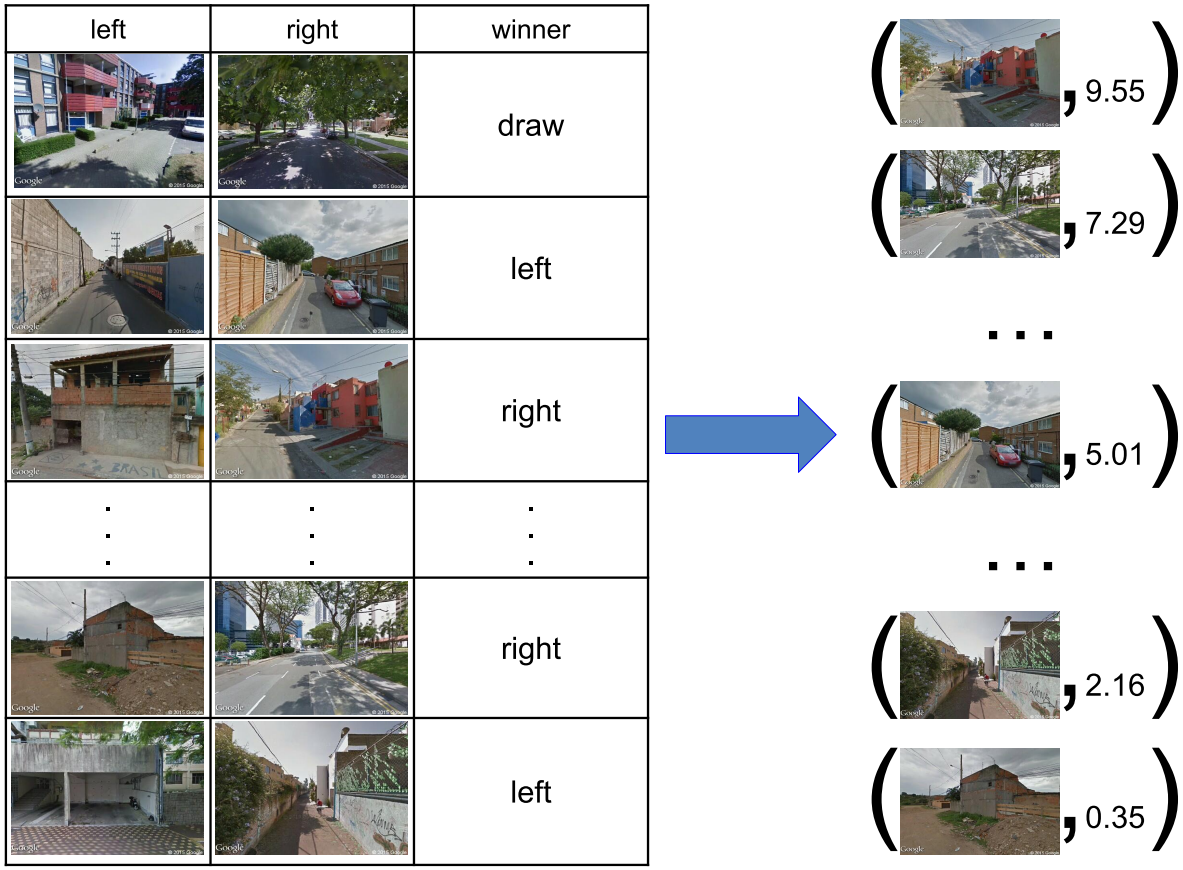

Figure 4 – Quantificando as avaliações humanas de percepção urbana utilizando SVI; a partir da comparação entre duas imagens, é possível calcular escores para cada imagem avaliada. Fonte: o autor.

podemos calcular taxas, prêmios e penalidades para cada imagem:

$$Rate_i^k = \frac{w_i^k}{w_i^k + d_i^k + l_i^k} \tag{2.19a}$$

$$Award_i^k = \frac{1}{w_i^k} \sum_{j=1}^{n_1} \frac{w_j^k}{w_j^k + d_j^k + l_j^k} \tag{2.19b}$$

$$Penalty_i^k = \frac{1}{l_i^k} \sum_{j=1}^{n_2} \frac{l_j^k}{w_j^k + d_j^k + l_j^k} \tag{2.19c}$$

$$Q_i^k = \frac{10}{3}(Rate_i^k + Award_i^k - Penalty_i^k + 1) \tag{2.19d}$$

Onde w_i^k, d_i^k e l_i^k denotam o número de vezes em que a imagem i foi escolhida como vencedora, empatada ou perdedora; n_1 e n_2 são as contagens de vitórias e derrotas; *Award* é a taxa média de vitória quando a imagem i venceu, e *Penalty* é a taxa média de derrota quando ela perdeu. Os **Q-scores** são então escalonados de 0 (muito inseguro) a 10 (muito seguro). A Figura 4 ilustra como as comparações pareadas podem ser transformadas em medidas quantitativas, permitindo a atribuição de escores ou ranqueamentos às imagens.

A formulação SOS apresentada aqui é a referência canônica para todos os cálculos

de escore de percepção nesta tese; os capítulos subsequentes remetem a esta seção em vez de derivá-la novamente de forma independente.

2.4.3 Tarefas de percepção

A modelagem da percepção urbana é predominantemente formulada como um problema de aprendizado supervisionado, em que os modelos são treinados para prever julgamentos humanos coletados por meio de pesquisas com base em SVI. Dependendo da natureza da variável-alvo e das anotações disponíveis, os estudos existentes normalmente formulam o problema de acordo com um de três paradigmas de aprendizado.

(i) Classificação trata a percepção urbana como uma tarefa de predição categórica, atribuindo rótulos discretos (por exemplo, *seguro* ou *inseguro*) às cenas urbanas. Dada uma cena urbana de entrada $\mathbf{x}$, um classificador prevê a distribuição de probabilidade sobre K classes de percepção,

$$p(y = k \mid \mathbf{x}) = \text{softmax}(f_\theta(\mathbf{x}))_k, \tag{2.20}$$

onde $f_\theta(\cdot)$ denota o modelo parametrizado por θ. O modelo é comumente otimizado usando a perda de entropia cruzada,

$$\mathcal{L}_{\text{cls}} = -\sum_{k=1}^{K} y_k \log p(y = k \mid \mathbf{x}), \tag{2.21}$$

onde y_k é o rótulo de referência codificado em one-hot. Os modelos de classificação frequentemente aproveitam características visuais extraídas de SVI e podem incorporar fontes complementares de informação, incluindo indicadores demográficos, atributos geoespaciais e descrições textuais, para melhorar o desempenho preditivo.

(ii) Regressão visa estimar escores perceptuais contínuos derivados de avaliações humanas agregadas. Em vez de prever categorias discretas, o modelo produz um escore escalar,

$$\hat{s} = f_\theta(\mathbf{x}), \tag{2.22}$$

onde $\hat{s} \in \mathbb{R}$ representa o escore de percepção previsto. O treinamento é tipicamente realizado minimizando o erro quadrático médio (MSE),

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{s}_i - s_i\right)^2, \tag{2.23}$$

onde s_i denota o escore de percepção de referência. Essa formulação permite uma representação mais matizada das percepções subjetivas ao capturar graus variados de atributos. As abordagens baseadas em regressão são particularmente úteis quando os

escores de percepção são medidos em escalas numéricas e predições de granularidade fina são desejadas.

(iii) Ranqueamento e aprendizado pareado foca em aprender preferências relativas entre cenas urbanas, em vez de rótulos ou escores absolutos. Dado um par de cenas urbanas $(\mathbf{x}_i, \mathbf{x}_j)$ com escores latentes correspondentes r_i e r_j, a probabilidade de que a cena i seja preferida à cena j pode ser modelada como

$$P(i \succ j) = \sigma(r_i - r_j) = \frac{1}{1 + \exp(-(r_i - r_j))}, \tag{2.24}$$

onde $\sigma(\cdot)$ é a função sigmoide. A perda de ranqueamento pareado correspondente é:

$$\mathcal{L}_{\text{rank}} = -\log P(i \succ j) = \log\left(1 + \exp(-(r_i - r_j))\right). \tag{2.25}$$

Os modelos são tipicamente treinados usando perdas de ranqueamento ou arcabouços probabilísticos, incluindo formulações baseadas em Elo e TrueSkill, para inferir ranqueamentos perceptuais latentes a partir de comparações pareadas.

Embora essas formulações difiram em seus objetivos de predição e sinais de supervisão, todas se apoiam no aprendizado de representações que capturam as características visuais e contextuais subjacentes à percepção humana dos ambientes urbanos.

2.4.4 Conjunto de dados de percepção

Os experimentos ao longo desta tese se baseiam em um único conjunto de dados de percepção em larga escala. Utilizamos o conjunto de dados Place Pulse 2.0, que, em sua versão original, consiste em aproximadamente 1,22 milhão de comparações pareadas brutas de imagens envolvendo 111.390 SVIs de 56 cidades em 28 países, coletadas usando o *Google Street View* (GSV). As imagens foram anotadas em seis categorias perceptuais: *segurança*, *vivacidade*, *beleza*, *riqueza*, *depressão* e *tédio*. Cada registro de comparação fornece os identificadores das imagens, as coordenadas geográficas e a identidade da imagem vencedora para a categoria perceptual correspondente. **O conjunto de dados divulgado não inclui informações por anotador, como nacionalidade ou dados demográficos; consequentemente, nenhuma característica pessoal dos anotadores (voluntários) pode ser analisada.**

Para transformar os resultados das comparações pareadas em escores perceptuais quantitativos, aplicamos o algoritmo SOS (Park and Newman, 2005) apresentado na Seção 2.4.2. Instanciando a formulação canônica (Equação (2.19d)) para a categoria *segurança* (k), o Q-score perceptual de cada imagem i é:

$$Q_i^k = \frac{10}{3}\left(\text{Rate}_i^k + \text{Award}_i^k - \text{Penalty}_i^k + 1\right) \tag{2.26}$$

onde Rate_i^k, Award_i^k e Penalty_i^k são definidos como na Seção 2.4.2, e o escore resultante é escalonado para o intervalo $[0, 10]$, onde um valor próximo de zero indica uma percepção muito insegura e um valor próximo de dez indica uma percepção muito segura. A Tabela 2 reporta as estatísticas do subconjunto retido após o pré-processamento das comparações.

Place Pulse 2.0		
Continent	# cities	# images
America	22	50,028
Europe	22	38,747
Asia	7	11,417
Oceania	2	6,097
Africa	3	5,101
Total	56	111,390

(a)

Place Pulse 2.0		
Category	# comparisons	*mean*
Safety	368,926	5.188
Lively	267,292	5.085
Beautiful	175,361	4.920
Wealthy	152,241	4.890
Depressing	132,467	4.816
Boring	127,362	4.810
Total	1,223,649	

(b)

Table 2 – Estatísticas obtidas após o processamento de todas as comparações do Place Pulse, contendo informações sobre imagens por cidade em cada continente (a) e o escore médio para cada categoria solicitada (b).

2.5 Considerações Finais

Este capítulo apresentou e descreveu os principais conceitos de aprendizado de máquina e aprendizado profundo comumente aplicados às análises de percepção urbana utilizando SVI. Começando pelos algoritmos clássicos de aprendizado de máquina e modelos lineares, e avançando até as arquiteturas de aprendizado profundo, cobrimos os principais paradigmas de aprendizado, as formulações de problemas e as tarefas relevantes para o processamento de SVI. Distinguimos ainda entre interpretabilidade e explicabilidade, revisando os métodos de atribuição e contrafactuais mais amplamente usados para compreender o comportamento do modelo. Em seguida, apresentamos o aprendizado de representação multimodal e os VLMs, desde o alinhamento contrastivo imagem-texto até as arquiteturas generativas e de seguimento de instruções, que fundamentam os componentes baseados em linguagem e de edição de imagem das contribuições apresentadas posteriormente nesta tese. Por fim, os conceitos de percepção urbana e os arcabouços de avaliação foram apresentados para contextualizar as contribuições apresentadas nos capítulos seguintes. Com esses fundamentos estabelecidos, o Capítulo 3 apresenta uma revisão sistemática da literatura sobre como esses métodos foram aplicados no campo da percepção urbana. A revisão abrange 234 estudos que abarcam abordagens de classificação, regressão, ranqueamento, baseadas em segmentação e multimodais para a percepção urbana, e conclui com quatro

lacunas metodológicas que motivam diretamente as contribuições empíricas e o projeto do sistema apresentados no restante desta tese.

3 Revisão Sistemática da Literatura

Como motivado no Capítulo 1, um desafio central na pesquisa em percepção urbana é que as abordagens existentes focam predominantemente na acurácia de predição, oferecendo pouca percepção sobre os fatores visuais e semânticos que impulsionam os julgamentos humanos de segurança urbana. Para caracterizar sistematicamente o estado do campo e identificar as lacunas metodológicas que esta tese aborda, conduzimos uma revisão estruturada da literatura sobre percepção urbana utilizando SVI.

Esta revisão sistemática segue o protocolo Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) (Regona et al., 2022)[1]. O PRISMA é um arcabouço amplamente reconhecido que aprimora a transparência, a consistência e a minúcia das revisões sistemáticas. Ele foi utilizado em estudos relacionados (Charreire et al., 2014; Yin et al., 2023; Marasinghe et al., 2023; He and Li, 2021). O método é dividido em quatro etapas: identificação, triagem, elegibilidade e inclusão. A revisão conclui identificando quatro lacunas metodológicas persistentes na literatura — percepção limitada sobre os fatores visuais, explicações estáticas e indiretas, edições generativas irrealistas e a ausência de ferramentas interativas com humano no processo (human-in-the-loop) — cada uma das quais é diretamente abordada por uma ou mais contribuições desta tese.

3.1 Metodologia

A Figura 5 ilustra esse processo de revisão. Primeiro, selecionamos palavras-chave relevantes para identificar um conjunto inicial de artigos, que foram triados para filtrar trabalhos irrelevantes. Em seguida, revisamos o conteúdo dos artigos restantes, focando em estudos relacionados à percepção urbana utilizando SVI de 2005 a 2026. O ano inicial foi selecionado devido ao lançamento das principais plataformas de *street view*, incluindo o Baidu Street View (2005), o Google Street View (2007) e o Tencent Street View (2011). Durante as etapas de elegibilidade e inclusão, empregamos o *snowballing* (Wohlin, 2014) para descobrir pesquisas relevantes adicionais não identificadas na busca inicial. Mais detalhes desse processo são descritos nas seções seguintes.

3.1.1 Identificação

Esta etapa, também conhecida como *critério de busca*, teve como objetivo identificar um conjunto inicial de artigos. Buscamos no ScienceDirect, Google/Semantic Scholar e Web of Science usando termos como *street view*, *street view imagery*, *street imagery* e *street*

[1] http://prisma-statement.org/

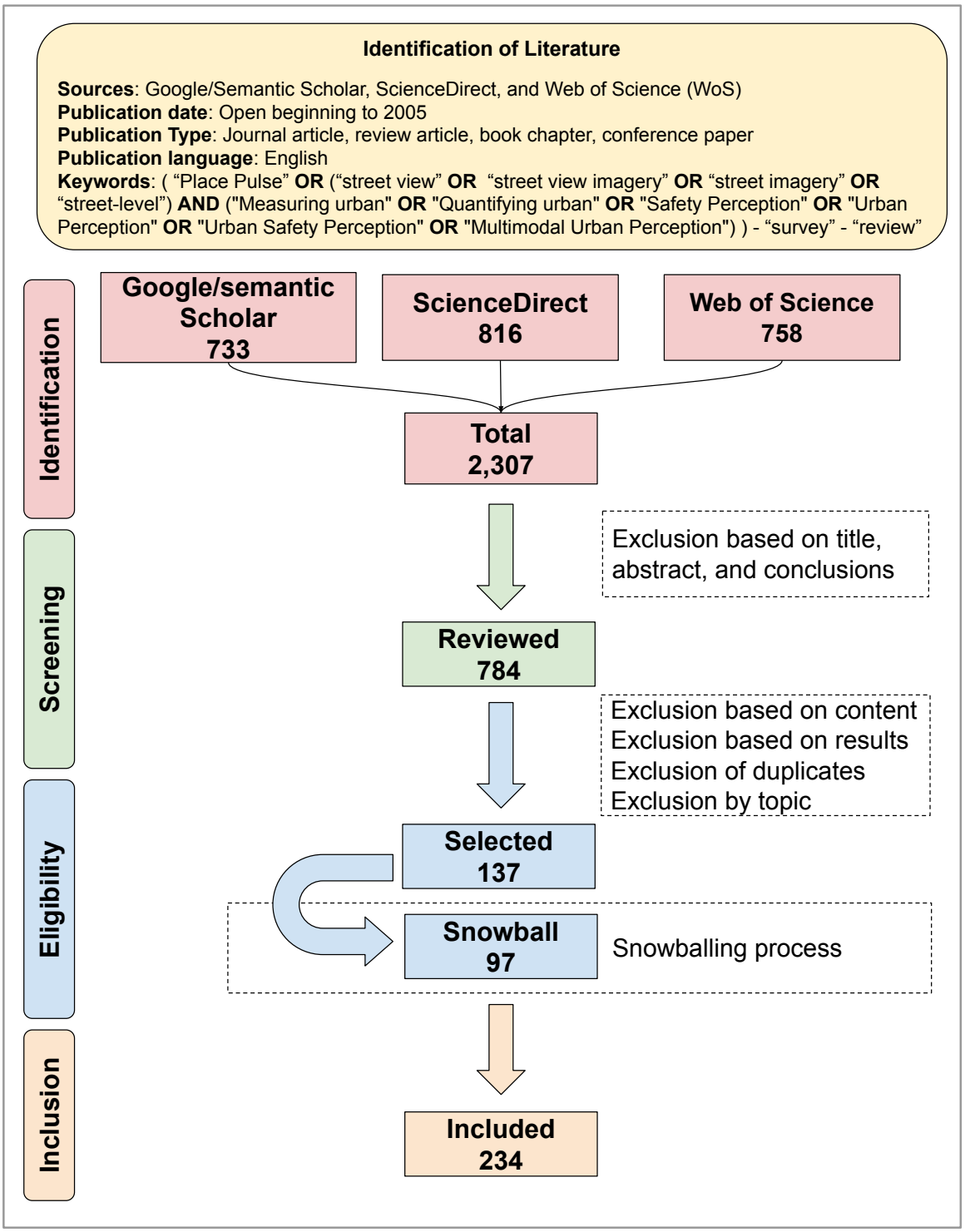

Figure 5 – Fluxo da revisão sistemática de acordo com o protocolo *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA). Observe que também incluímos *snowballing* na última etapa do subprocesso para revisar e capturar mais artigos para esta revisão. Fonte: o autor.

level. Palavras-chave adicionais, incluindo *Measuring urban*, *Quantifying urban*, *Safety perception*, *Urban Perception*, *Multimodal Urban Perception* e *Urban Safety Perception*, foram usadas para capturar estudos sobre percepção urbana utilizando SVI. O termo

Table 3 – Número de artigos por etapa e palavras-chave

Search Keywords	Identification	Screening	Eligibility	Snowballing	Final Inclusion
Street View	1,090	312	17	12	29
Measuring urban	635	118	9	5	14
Quantifying urban	252	74	13	5	18
Urban perception	184	156	39	58	97
Place-Pulse	28	28	28	3	31
Safety perception	86	64	15	8	23
Multimodal urban perception	32	32	16	6	22
Total	2,307	784	137	97	234

Place-Pulse também foi incluído para identificar estudos que seguiram a metodologia daquele conjunto de dados. A Tabela 3 mostra o número de artigos identificados e retidos ao longo do processo de revisão após a remoção de duplicatas.

Embora termos como *street view* e *street* sejam genéricos, eles ajudam a recuperar artigos que referenciam serviços como *Google Street View*, *Tencent Street View*, *Naver Street View*, *Baidu Street View* e *OpenStreetMap*, entre outros serviços. A busca foi expandida para incluir serviços de mapeamento relacionados: combinar *mapping*, *street view* e *urban perception* resultou em 39 resultados, enquanto excluir *urban perception* produziu 1.470. Estudos irrelevantes (por exemplo, poluição, inferência térmica, sensoriamento remoto, ambientes aéreos, mudança temporal e análise meteorológica) foram excluídos. Embora *Street View* tenha retornado mais resultados, a maioria foi excluída; *Urban perception* e *Place-Pulse* produziram os estudos mais relevantes. No total, esta etapa identificou **2.307 artigos**.

3.1.2 Triagem

Esta etapa, também chamada de *critério de seleção*, envolveu revisar os títulos, resumos, dados, métodos e conclusões dos 2.307 artigos iniciais para identificar um subconjunto relevante. Os critérios de inclusão foram: (1) idioma inglês; (2) foco na percepção urbana em um contexto em nível de rua; (3) ir além da visão computacional ao incorporar abordagens como análise estatística e aprendizado profundo; e (4) examinar as preferências de percepção humana coletadas por meio de SVI e pesquisas online. **Excluímos** tópicos como clima e meteorologia, poluição, estudos populacionais, análise solar e térmica, monitoramento agrícola, tráfego e simulação de cidades, bem como estudos focados exclusivamente em tarefas de visão computacional ou aprendizado profundo (por exemplo, modelos de segmentação de imagens em cenas urbanas). Esta etapa resultou em um total de **784 artigos**.

3.1.3 Elegibilidade

Revisamos 784 artigos extraindo características-chave, incluindo serviço de SVI, ano de publicação, palavras-chave, veículo de publicação, afiliações dos autores, conjuntos de dados, localização do estudo, tarefa de aprendizado de máquina, referências e citações. Os artigos foram excluídos se fossem (1) duplicatas, (2) não revisados por pares (por exemplo, resumos, apresentações, livros, relatos de caso, preprints do arXiv ou similares) ou (3) não relacionados à percepção urbana. Esse processo resultou em **137 artigos**. Em seguida, aplicamos o *snowballing* para identificar estudos relevantes adicionais.

Processo de snowballing

O *snowballing*, ou encadeamento de citações, expande a literatura relevante para além dos resultados de busca iniciais Wohlin (2014). Ele envolve examinar iterativamente as referências dos artigos selecionados e identificar novos estudos relevantes por meio de etapas *para trás* (*backward*) e *para frente* (*forward*):

- **Snowballing para trás** parte das referências dos artigos identificados e trabalha retrospectivamente.
- **Snowballing para frente** explora os artigos que citaram os artigos identificados, avançando no tempo.

Usamos o Semantic Scholar[2] para identificar referências e citações altamente influentes, com base em sinais contextuais e na frequência de citação. Em ambas as etapas, também verificamos o idioma, a origem geográfica e o tipo de veículo, seguindo os mesmos critérios de filtragem da etapa de triagem anterior. Aplicando isso aos 137 artigos iniciais, selecionamos 50 artigos adicionais na etapa para trás e 47 na etapa para frente, resultando em um total de **234 artigos incluídos nesta revisão**.

3.1.4 Inclusão final

De um conjunto inicial de 2.307 artigos, 234 atenderam aos critérios de inclusão. Esses estudos foram sistematicamente analisados e categorizados usando um arcabouço multimodal de quatro componentes que reflete o típico *pipeline* orientado a dados na pesquisa em percepção urbana, desde a coleta e integração de dados até a análise e a aplicação no mundo real, ao mesmo tempo em que incorpora uma abordagem multimodal. Dada a sua amplitude e diversidade, acreditamos que esta revisão captura as principais tendências, métodos e direções emergentes, minimizando o viés de seleção. A lista completa dos artigos incluídos está disponível em um repositório público.[3]

[2] https://www.semanticscholar.org/

[3] https://docs.google.com/spreadsheets/d/1P_OTmYEz8dKRphqIe8PLh_JouPrMlGAclnALWfohSts/

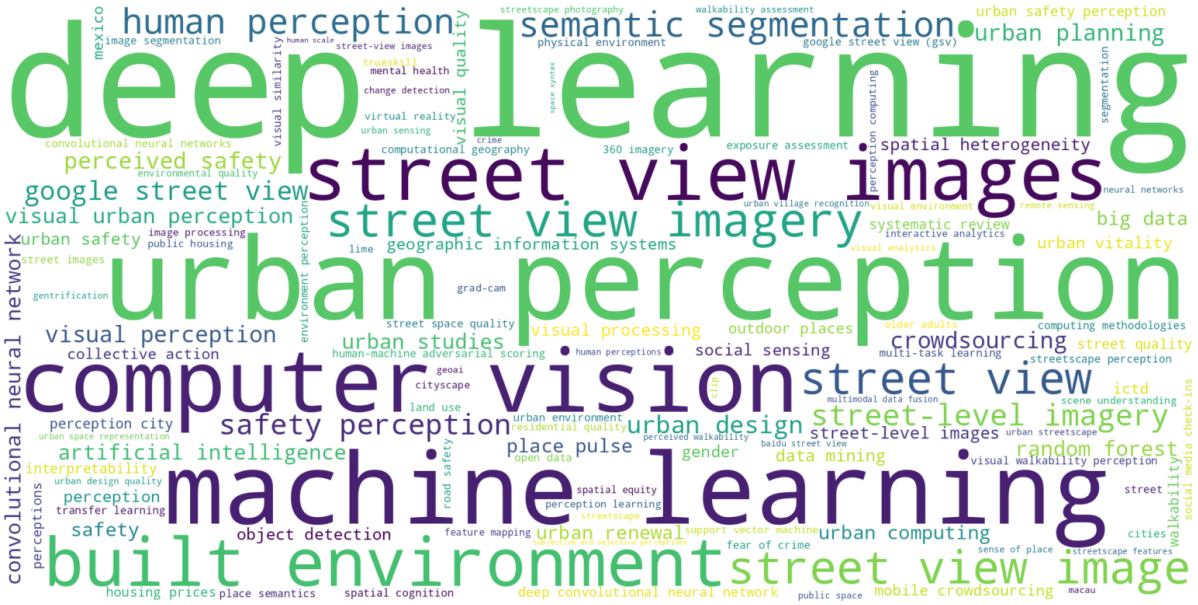

Figure 6 – Nuvem de palavras que retrata as palavras-chave mais frequentes dos 234 estudos incluídos. Fonte: o autor.

3.2 Resultados descritivos

Aqui, apresentamos os principais achados descritivos obtidos a partir da *Systematic Literature Review* (SLR).

3.2.1 Visão geral do corpus

Os 234 estudos selecionados foram publicados em 113 veículos distintos, ressaltando o caráter inerentemente interdisciplinar da pesquisa em percepção urbana. Os periódicos com maior representatividade são *Cities* (16 artigos), *Landscape and Urban Planning* (12 artigos), *Computers, Environment, and Urban Systems* (11 artigos) e *ISPRS International Journal of Geo-Information* (10 artigos). Essa distribuição ilustra a interseção de vários domínios de pesquisa, incluindo estudos urbanos, ciência da informação geográfica, visão computacional, sensoriamento remoto e inteligência artificial.

A análise de palavras-chave destaca ainda mais a orientação metodológica do campo. Os termos mais frequentemente reportados incluem *deep learning*, *urban perception*, *machine learning*, *computer vision* e *street view images*. Em conjunto, essas palavras-chave indicam uma forte dependência de representações baseadas em imagens e de técnicas orientadas por IA para analisar e modelar as percepções humanas dos ambientes urbanos. A distribuição completa de frequência das palavras-chave é apresentada na Figura 6.

3.2.2 Tendências de publicação

Embora o período de busca tenha começado em 2005, todos os estudos incluídos nesta revisão foram publicados após 2012. Esse período coincide tanto com a ascensão das redes neurais convolucionais (CNNs) quanto com os primeiros esforços em larga escala para quantificar a percepção urbana por meio de plataformas de *crowdsourcing* online Salesses et al. (2013), marcando o surgimento da pesquisa em percepção urbana baseada em imagens de *street view* como um campo distinto. As primeiras contribuições (2012–2014) empregavam predominantemente descritores visuais construídos manualmente, incluindo características Gist, histogramas de cor e representações baseadas em textura Doersch et al. (2012); Wilson et al. (2012); Lindal and Hartig (2013); Ordonez and Berg (2014); Naik et al. (2014). À medida que o aprendizado profundo amadureceu, essas abordagens foram progressivamente substituídas por arquiteturas baseadas em CNN e, em alguns casos, por técnicas de aprendizado por reforço Wang et al. (2022a,b).

Mais recentemente, o campo passou por uma nova mudança metodológica impulsionada pelo surgimento dos *Large Language Models* (LLMs) Zhang et al. (2025a); Malekzadeh et al. (2024); Huang et al. (2024); Levering et al. (2024) e dos modelos multimodais (principalmente VLMs) Moreno-Vera and Poco (2025); Zhang et al. (2025b); Han and Kim (2025), que possibilitam um raciocínio multimodal mais rico e representações mais flexíveis das cenas urbanas.

A Figura 7 reporta as contagens anuais de publicações juntamente com os totais cumulativos de citações por ano de publicação, obtidos por meio da API do Semantic Scholar. A atividade de publicação permaneceu relativamente estável durante 2014–2016 e 2018–2020, mas acelerou consideravelmente após 2020. Esse crescimento parece estar associado à crescente disponibilidade de conjuntos de dados de SVI em larga escala, à incorporação de fontes complementares de dados urbanos e à ampla adoção de modelos de fundação e técnicas de aprendizado multimodal.

É importante notar que os valores de citação correspondem ao número total de citações acumuladas pelos artigos publicados em um dado ano, e não às citações recebidas durante aquele ano. Consequentemente, publicações recentes naturalmente exibem contagens de citações mais baixas devido ao seu menor tempo de exposição. Até abril de 2026, os estudos incluídos nesta revisão haviam acumulado coletivamente 16.967 citações. Cinco publicações de 2026 também foram incluídas, o que explica a ausência de citações para aquele ano.

3.2.3 Geografia

A análise das afiliações dos autores revela uma forte concentração da atividade de pesquisa em um pequeno número de países. A China responde por 35,43% de todas as afiliações, seguida pelos Estados Unidos (17,21%) e pelo Reino Unido (8,60%). No entanto,

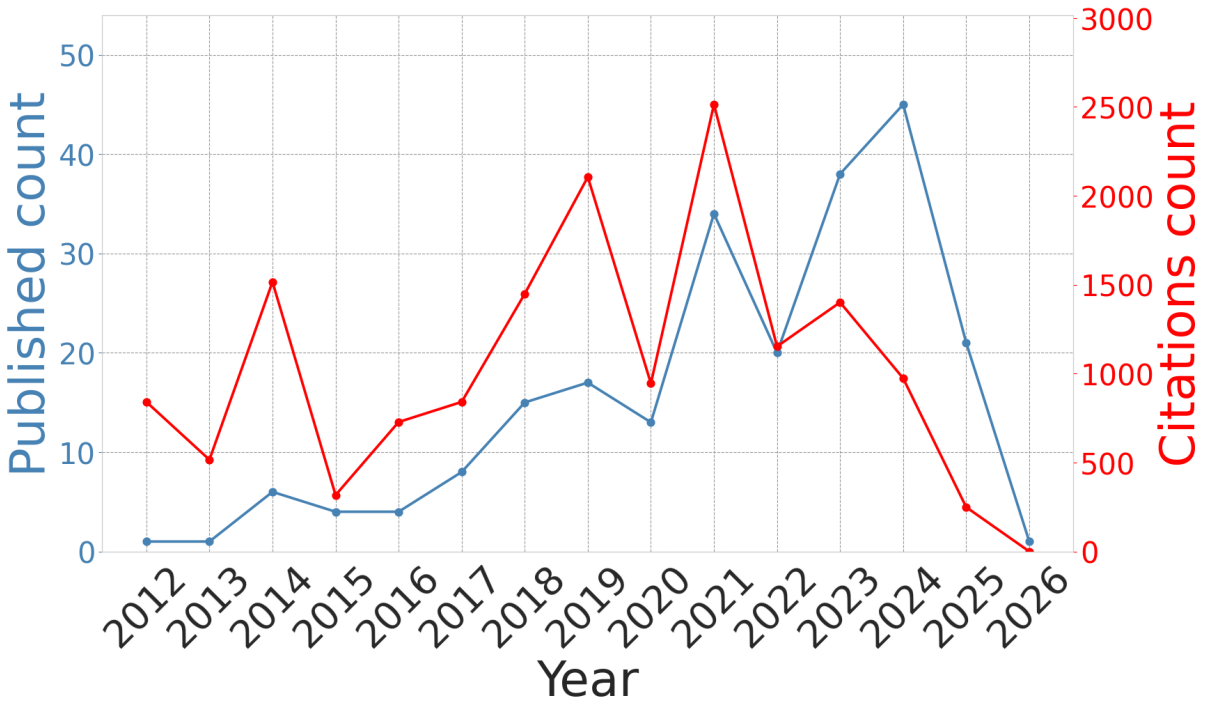

Figure 7 – O número de artigos publicados (azul) e de citações (vermelho) por ano mostra um crescimento claro na produção de pesquisa sobre percepção urbana utilizando SVI, particularmente após 2016, com a contagem de publicações atingindo seu pico entre 2021 e 2024. As citações exibem um padrão mais flutuante, alcançando seus níveis mais altos por volta de 2019 e 2021, seguidos por um declínio notável nos últimos anos. A queda acentuada nas citações para 2024–2026 é esperada, uma vez que essas publicações são mais recentes e tiveram menos tempo para acumular citações. Fonte: o autor.

a ênfase temática da pesquisa em percepção urbana varia entre as regiões.

A pesquisa conduzida por instituições sediadas nos EUA foca principalmente em melhorar os modelos preditivos e desenvolver novos métodos computacionais para a estimativa da percepção. Em contraste, as instituições chinesas frequentemente enfatizam tanto o desenvolvimento metodológico quanto aplicações em larga escala a cidades em rápida urbanização, muitas vezes adaptando ou estendendo conjuntos de dados de referência (benchmark) como o Place Pulse 2.0 Dubey et al. (2016). Os estudos sediados no Reino Unido comumente investigam as relações entre a forma urbana, a estética e a percepção humana, incluindo avaliações de atratividade visual e qualidade ambiental Seresinhe et al. (2017); Law et al. (2018b,a); Liu et al. (2023b); Muller et al. (2022); Joglekar et al. (2020); Quercia et al. (2014).

O crescimento da produção de pesquisa chinesa entre 2016 e 2019, visível na Figura 7, coincide com a adoção mais ampla de técnicas de aprendizado profundo e a crescente acessibilidade de conjuntos de dados de *street view* chineses em larga escala.

Embora a produção de pesquisa seja dominada pela China, pelos Estados Unidos e pelo Reino Unido, a literatura revisada abrange 41 países adicionais. Quanto às localizações

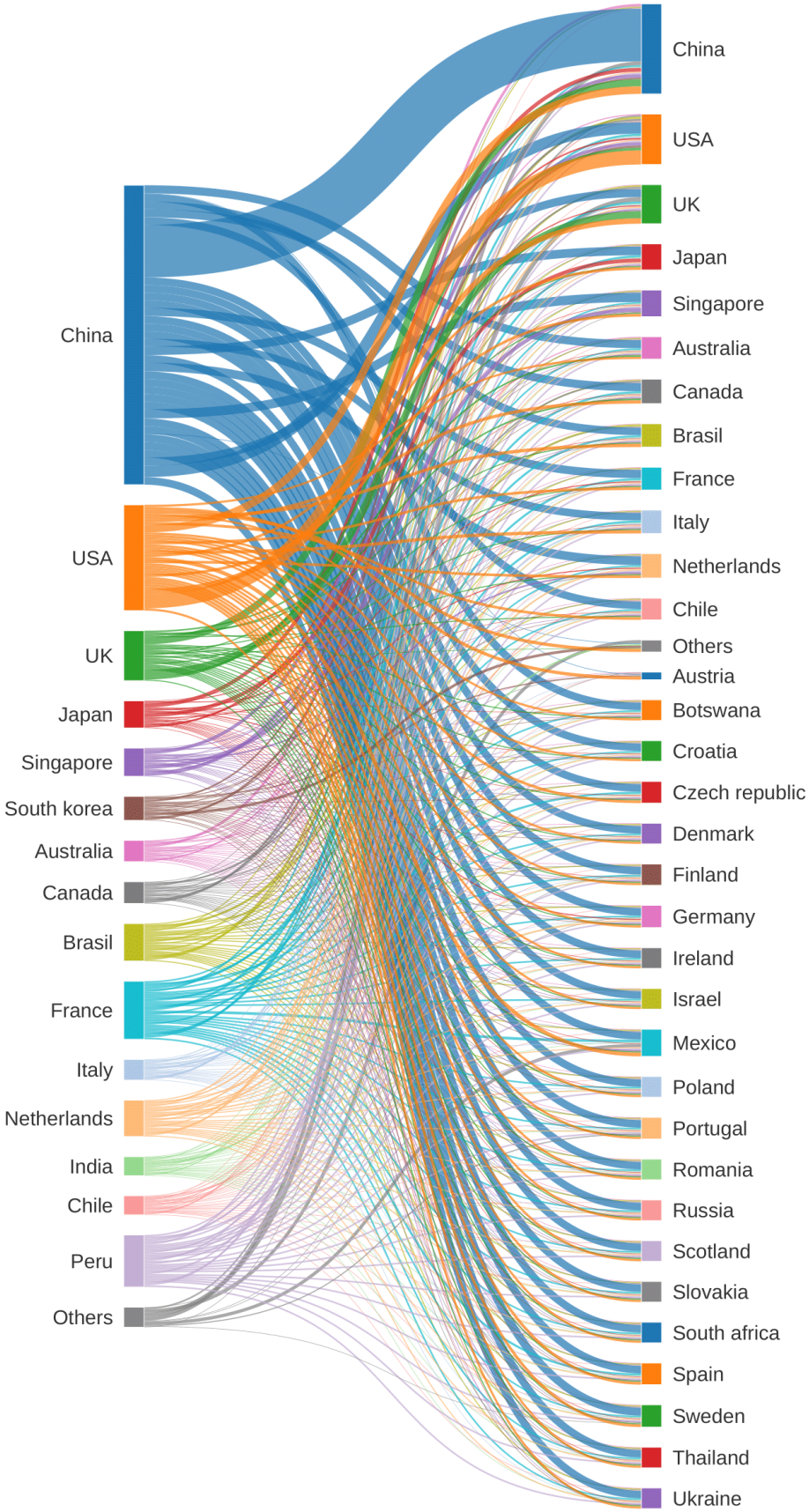

Figure 8 – A relação entre as afiliações dos autores, o país (à esquerda) e os países dos estudos de aplicação (à direita). Fonte: o autor.

Table 4 – Provedores de serviço de SVI mais utilizados e o número de estudos em diferentes escalas geográficas. Os estudos podem aparecer em múltiplas categorias de escala geográfica.

Provider	Resolution (max)	Global	Country	City	Total
GSV	2048×2048	50	1	85	136
BSV	1024×512	7	3	58	68
TSV	1680×1200	5	0	19	24
NSV	1200×628	0	0	4	4
Others	Depends on provider	24	11	92	127

dos estudos, a China (10,35%), os Estados Unidos (5,89%) e o Reino Unido (4,28%) também são os cenários mais frequentemente investigados. Os percentuais de afiliação e de país do estudo foram computados em nível de artigo, usando as afiliações institucionais de todos os autores e o foco geográfico de cada estudo. A Figura 8 ilustra essas relações por meio de um diagrama de Sankey, destacando a China como a localização de estudo mais proeminente em todo o corpus.

3.2.4 Provedores

As imagens de *Street View* (SVI) tornaram-se uma fonte de dados fundamental para a pesquisa em percepção urbana porque possibilitam a observação virtual em larga escala de paisagens urbanas e ambientes urbanos Li et al. (2022b). Em muitos estudos, as redes viárias são primeiro obtidas de fontes como o OpenStreetMap (OSM), o GeoGraph ou plataformas GIS antes que as imagens sejam sistematicamente coletadas dos serviços de *street view*. Entre os provedores identificados nesta revisão, o Google Street View (GSV), o Tencent Street View (TSV) e o Baidu Street View (BSV) são de longe os mais amplamente utilizados. O GSV oferece a mais ampla cobertura internacional, atualmente abrangendo mais de 135 países, enquanto o BSV e o TSV fornecem funcionalidade semelhante com um foco mais forte em cidades chinesas, cobrindo aproximadamente 652 e 296 cidades, respectivamente.

Várias alternativas regionais também aparecem na literatura, incluindo o Naver Street View (NSV) e o Kakao Street View (KSV) na Coreia do Sul, bem como o CycloMedia em partes da Europa. Além disso, plataformas orientadas pela comunidade, como o Mapillary e o KartaView, fornecem extensos repositórios de imagens que, coletivamente, contêm aproximadamente 1,5 bilhão de imagens, embora a cobertura e a qualidade dependam em grande parte das contribuições dos usuários. Para caracterizar melhor as práticas de coleta de dados, classificamos ainda os estudos de acordo com seu escopo geográfico: estudos em *escala de cidade*, que focam em uma única cidade; estudos em *escala de país*, que abrangem múltiplas cidades dentro de uma nação; e estudos em *escala global*, que abrangem múltiplos países. A Tabela 4 resume os principais serviços de SVI usados ao

longo da literatura. Como vários estudos empregam mais de um provedor de imagens e operam em múltiplas escalas geográficas, as contagens reportadas são não exclusivas.

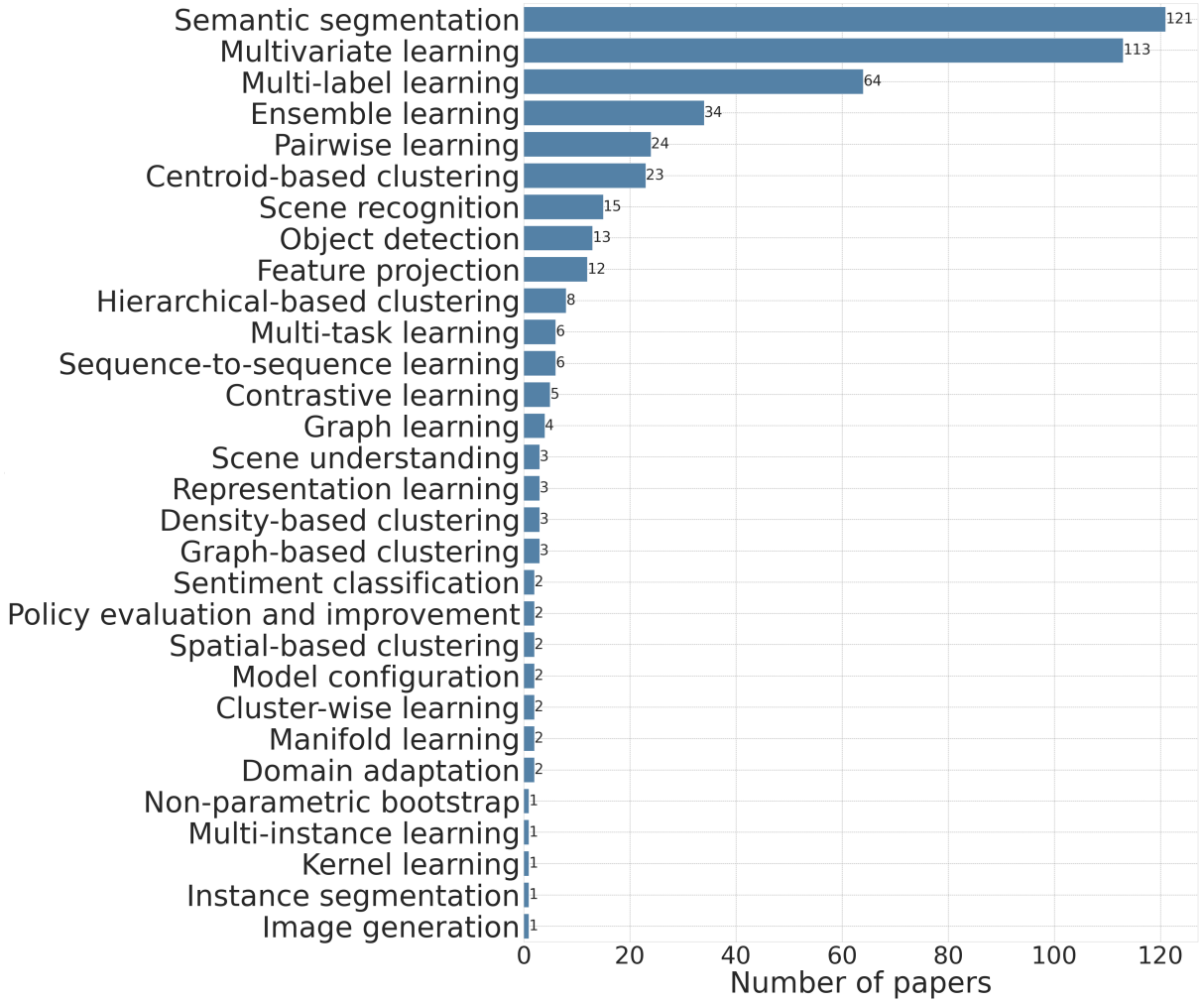

Figure 9 – O número de estudos por categoria de tarefa mostra que a segmentação semântica e o aprendizado multivariado são os métodos mais comuns, seguidos pelo aprendizado *multi-label* e *ensemble*. Observamos que alguns estudos realizam múltiplas tarefas simultaneamente, o que os coloca em múltiplas categorias.

3.2.5 Problemas e tarefas de aprendizado

Apoiando-se na taxonomia de aprendizado introduzida no Capítulo 2 (tipos, problemas e tarefas de aprendizado), a Tabela 5 resume as tarefas de aprendizado identificadas nos estudos selecionados, agrupadas por paradigma. O aprendizado supervisionado domina, liderado pela classificação (176 estudos, 43,24%) e pela regressão (134, 32,92%), refletindo a prevalência tanto de rótulos categóricos quanto de escores de percepção contínuos nos conjuntos de dados de percepção urbana. O aprendizado não supervisionado inclui o agrupamento (46, 11,30%) e a redução de dimensionalidade (14, 3,44%) — esta última referindo-se a técnicas que comprimem representações de características de SVI de alta dimensionalidade em espaços de menor dimensionalidade para visualização ou modela-

gem posterior, complementando o problema de agrupamento definido no Capítulo 2. As abordagens autossupervisionadas — aqui abrangendo modelos generativos e VLMs, que aprendem representações sem rótulos explicitamente fornecidos por humanos — compreendem os modelos de linguagem multimodais (16, 3,93%), que começaram a surgir em 2024, e os modelos generativos (6, 1,47%). O aprendizado por reforço, apesar de receber cobertura dedicada no Capítulo 2 devido à sua relevância conceitual, aparece em apenas 2 estudos (0,49%) dentro do corpus de percepção urbana com SVI, refletindo sua limitada adoção prática nesse domínio até o momento. Os percentuais são computados sobre as 407 ocorrências totais de tarefas, uma vez que muitos artigos empregaram múltiplas tarefas simultaneamente.

Table 5 – Tipo de aprendizado, problema de aprendizado e número de estudos.

Learning type	Learning problem	Studies
Supervised Learning	Classification	176
	Regression	134
	Learning to rank	11
	Multitask	8
	Density estimation	2
Unsupervised Learning	Clustering	46
	Dimensionality reduction	14
Self-Supervised Learning	Multimodal Language models	16
	Generative models	6
Reinforcement Learning	Policy learning	2
Total approaches		407

Entre as tarefas específicas, a segmentação semântica e o aprendizado multivariado são as mais comuns (Figura 9), com muitos estudos analisando razões de pixels em nível de objeto e a importância dos objetos. Tarefas raras — aparecendo em no máximo um estudo — incluem *style transfer*, *policy learning*, *graph-based clustering*, *depth estimation* e *kernel learning*.

3.2.6 O papel emergente dos VLMs

Uma tendência emergente em toda a literatura revisada é o uso crescente de VLMs e LLMs na pesquisa em percepção urbana. Além de sua aplicação para avaliar, descrever e interpretar cenas de *street view*, estudos recentes têm aproveitado esses modelos para tarefas como edição de imagens, geração de cenários e análise contrafactual, permitindo que os pesquisadores explorem como modificações nos ambientes urbanos podem influenciar as percepções e experiências humanas Huang et al. (2024); Moreno-Vera and Poco (2025); Zhang et al. (2025a); Malekzadeh et al. (2024); Han and Kim (2025); Zhang et al. (2025b); Levering et al. (2024); Law et al. (2023); Hu et al. (2023); Tan et al. (2025); Hu et al. (2025). Essa tendência reflete uma mudança mais ampla em direção a abordagens multimodais que integram conjuntamente informação visual e linguística, permitindo representações

mais ricas dos ambientes urbanos e arcabouços mais flexíveis para compreender, prever e moldar as percepções humanas das cidades. Essa trajetória motiva diretamente as escolhas metodológicas adotadas na presente tese, que aproveita os VLMs tanto para a avaliação da percepção quanto para a edição contrafactual fundamentada de cenas.

3.3 Considerações finais

Este capítulo apresentou uma revisão sistemática da literatura sobre pesquisa em percepção urbana utilizando SVI, seguindo o protocolo PRISMA. Por meio de um processo estruturado de identificação, triagem e filtragem por elegibilidade, um conjunto inicial de 2.307 artigos foi progressivamente reduzido a 137 estudos. A aplicação do *snowballing* para frente e para trás rendeu 50 e 47 artigos adicionais, respectivamente, resultando em um corpus final de **234 estudos**. A análise descritiva desse corpus revelou padrões-chave em tendências de publicação, distribuição geográfica, provedores de SVI e abordagens metodológicas. Notavelmente, a produção de pesquisa acelerou consideravelmente desde 2020, impulsionada pela adoção de modelos de fundação e técnicas de aprendizado multimodal. Uma tendência emergente nos estudos recentes é o uso crescente de VLMs para tarefas de percepção urbana, sinalizando uma mudança mais ampla em direção a representações mais ricas e fundamentadas na linguagem dos ambientes urbanos.

A revisão também revelou um conjunto de lacunas de pesquisa persistentes que motivam as contribuições desta tese:

- **Percepção limitada sobre os fatores visuais**: a maioria dos estudos concentra-se em melhorar a acurácia de predição, fornecendo pouca explicação sobre quais elementos visuais e semânticos específicos impulsionam os julgamentos humanos de segurança. *Abordado no Capítulo 4.*
- **Explicações estáticas e indiretas**: quando métodos de interpretabilidade são aplicados, as saídas resultantes — *saliency maps* ou escores de atribuição computados sobre representações abstratas de características — estão frequentemente desconectadas do arranjo espacial da própria cena e de modificações de cena orientadas ao planejamento. *Abordado conjuntamente nos Capítulos 4 e 5: o primeiro fundamenta a explicação no arranjo espacial da própria cena, e o segundo traduz as diferenças resultantes em recomendações legíveis por humanos.*
- **Edições generativas irrealistas**: as abordagens de edição de imagens contrafactuais e generativas tendem a otimizar as edições para alterar a saída de um modelo sem fundamentar as transformações em diferenças observadas entre cenas urbanas reais, limitando sua credibilidade como evidência para decisões de projeto. *Abordado nos Capítulos 6 e 7.*

- **Ausência de ferramentas interativas com humano no processo (human-in-the-loop)**: nenhum sistema anterior integra modelagem preditiva, geração contrafactual fundamentada e análise visual interativa em um único ambiente exploratório adaptado à percepção urbana. *Abordado nos Capítulos 6 e 7.*

Essas quatro lacunas mapeiam diretamente para as três contribuições empíricas que se seguem. O Capítulo 4 retoma a primeira lacuna e começa a fechar a segunda ao fundamentar a explicação no arranjo espacial da própria cena, em vez de em um vetor de características abstraído. O Capítulo 5 completa a segunda lacuna ao traduzir essas diferenças espacialmente fundamentadas em recomendações legíveis por humanos. Os Capítulos 6 e 7 então fecham a terceira e a quarta lacunas em conjunto: o UrbanCounterViz posiciona-se em relação aos sistemas anteriores específicos em análise visual, modelagem computacional da percepção e edição generativa de cenas mais próximos a ele, fundamenta cada edição de cena em transformações urbanas reais e observadas, em vez de geração aberta, e incorpora todo o *pipeline* em um ambiente interativo, com humano no processo (human-in-the-loop).

4 Identificando Elementos Visuais Relevantes na Percepção de Segurança Urbana

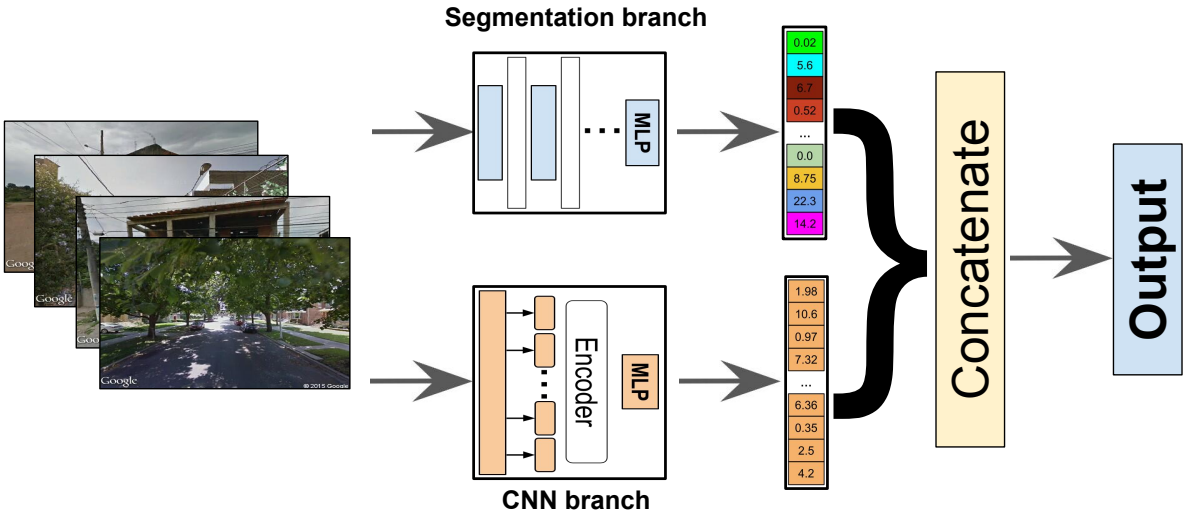

Figure 10 – O UrbanFormer proposto combina a saída vetorizada do OneFormer com uma saída de vetor de probabilidade do ConvNext modificado para prever a percepção humana da rua entre segura e insegura. Fonte: o autor.

4.1 Introdução

Este capítulo apresenta a **primeira contribuição empírica** desta tese (Contribuição 2, conforme enumerada no Capítulo 1): uma investigação sobre quais elementos visuais e semânticos mais fortemente influenciam as percepções humanas de segurança urbana. Como discutido no Capítulo 3, a maioria dos estudos de percepção urbana otimiza para a acurácia preditiva — reportando o quão bem um modelo classifica uma rua como segura ou insegura — oferecendo comparativamente pouca percepção sobre *por que* uma cena é percebida daquela forma. Modelos treinados diretamente em imagens brutas, em particular, tendem a se comportar como preditores opacos: podem alcançar um forte desempenho de classificação enquanto deixam o raciocínio visual por trás de suas decisões em grande parte inacessível ao analista.

Este capítulo aborda essa lacuna diretamente, mas a própria lacuna tem duas faces distintas na literatura anterior. Modelos treinados de ponta a ponta em imagens brutas alcançam um forte desempenho de classificação, mas raramente são submetidos a uma análise profunda, em nível de elemento, sobre aquilo em que a própria atenção da rede está

focada. As abordagens baseadas em segmentação, por outro lado, de fato oferecem uma forma de interpretabilidade — mas tipicamente apenas ao tratar a saída da segmentação como um vetor de características estático de razão de pixels e aplicar um método *post-hoc*, como LIME ou SHAP, para estimar a importância de cada categoria dentro desse vetor. Nenhuma das rotas olha de volta para a própria imagem: a primeira porque carece de um vocabulário semântico para descrever aquilo a que atende; a segunda porque sua explicação opera sobre um vetor abstraído, em vez de sobre as regiões espaciais da cena. Este capítulo propõe o **UrbanFormer**, um classificador que funde características visuais de CNN com um vetor explícito de composição semântica, combinado a um método de explicação que fecha essa lacuna remanescente: em vez de estimar a importância das características sobre o vetor de razão de pixels, ele aplica o Grad-CAM ao ramo CNN e interseca os mapas de atenção resultantes diretamente com as máscaras de segmentação, medindo a relevância dos elementos visuais como sobreposição de área espacial na própria imagem. O trabalho foi publicado na *IEEE/WIC/ACM International Conference on Web Intelligence* (WI-IAT '24) (Moreno-Vera et al., 2024). Todos os materiais podem ser encontrados aqui[1].

Contribuições. Este capítulo contribui com: (i) o UrbanFormer, um modelo que combina vetores de composição semântica derivados do OneFormer com características visuais de CNN, alcançando o mais forte desempenho reportado de classificação binária entre os modelos avaliados e um desempenho competitivo na tarefa de granularidade mais fina de 10 rótulos; e (ii) uma metodologia de explicação que cruza as ativações do Grad-CAM com as máscaras de segmentação para quantificar, elemento por elemento, quais partes de uma cena de rua estão mais associadas à percepção segura ou insegura — fornecendo o vocabulário visual sobre o qual os Capítulos 5 e 6 se apoiam diretamente.

4.2 Trabalhos Relacionados

As primeiras abordagens à percepção urbana computacional se apoiavam em descritores de imagem construídos manualmente — incluindo características GIST, ativações DeCAF e representações baseadas em ImageNet — para modelar correlações entre padrões visuais e julgamentos humanos (Naik et al., 2014; Ordonez and Berg, 2014; Naik et al., 2016). Esses métodos estabeleceram os pipelines fundamentais ainda usados hoje para coletar anotações perceptuais em larga escala por meio de *crowdsourcing*, mais notavelmente o conjunto de dados MIT Place Pulse (Salesses et al., 2013), e demonstraram que atributos perceptuais como segurança, beleza e vivacidade poderiam ser estimados a partir de dados visuais em escala. No entanto, a dependência de características projetadas manualmente limitou a capacidade representacional desses modelos e sua habilidade de capturar relações semânticas e contextuais mais ricas presentes nas cenas urbanas.

[1] https://visualdslab.com/papers/UrbanFormer/

A adoção subsequente de redes neurais convolucionais melhorou substancialmente o desempenho preditivo ao aprender características hierárquicas diretamente das imagens brutas, permitindo que os modelos capturassem aspectos de granularidade mais fina de textura, arranjo e contexto espacial (Dubey et al., 2016; Li et al., 2021; Zhang et al., 2021; Liu et al., 2022a; Zhou et al., 2019), e foram subsequentemente aplicadas a questões mais direcionadas, como a influência da vegetação e da desordem física na segurança percebida (Li et al., 2015a; Porzi et al., 2015; Arietta et al., 2014; Tokuda et al., 2019; Lavi. et al., 2022). Esses estudos estabeleceram *o que* pode ser previsto a partir de uma SVI, mas raramente examinaram quais elementos visuais específicos impulsionaram uma dada predição.

Mais recentemente, as abordagens baseadas em segmentação melhoraram a interpretabilidade ao representar as cenas urbanas em termos de categorias de objetos semanticamente significativas, em vez de pixels brutos, e ao relacionar a composição da cena em nível de pixel à segurança percebida (Zhang et al., 2018b; Min et al., 2020; Yao and Zhu, 2025; Zhou et al., 2025a; Liang et al., 2023; Zhu et al., 2024). Esses métodos fornecem uma base mais transparente para a modelagem da percepção do que as CNNs baseadas apenas em aparência, mas tipicamente usam a segmentação apenas como um auxílio de pré-processamento ou visualização, em vez de integrá-la diretamente ao próprio modelo preditivo.

Diferenciação. O trabalho anterior nesse espaço segue um de dois caminhos, e nenhum realiza o tipo específico de análise que este capítulo introduz. As CNNs baseadas em aparência alcançam um forte desempenho preditivo, mas raramente são submetidas a uma explicação profunda, em nível de elemento, sobre aquilo a que a rede atende. As abordagens baseadas em segmentação de fato fornecem uma forma de interpretabilidade, mas tipicamente ao aplicar um método *post-hoc*, como LIME ou SHAP, ao vetor de razão de pixels derivado da segmentação (Zhang et al., 2018b; Min et al., 2020) — estimando a importância de cada categoria semântica como uma característica abstrata, sem relacionar essa importância de volta a onde essas categorias de fato aparecem na imagem. O UrbanFormer se distingue de ambas: ele funde características visuais de CNN com o vetor de composição semântica do OneFormer para a classificação e, então, explica suas predições aplicando o Grad-CAM diretamente ao ramo CNN e intersecando os mapas de atenção resultantes com as próprias máscaras de segmentação. A relevância dos elementos visuais é assim medida como a sobreposição de área espacial entre aquilo a que a rede atende e onde cada elemento fisicamente aparece na cena — uma forma de explicação fundamentada na imagem, e não em um vetor de características abstraído. Esse projeto conjunto é o que permite ao capítulo responder não apenas *quão segura* uma cena é prevista ser, mas *quais elementos* da cena a predição pode ser rastreada até.

4.3 Metodologia

A metodologia está organizada em três etapas principais. A primeira etapa envolve a **quantificação da percepção urbana**: uma análise exploratória do conjunto de dados Place Pulse 2.0 para derivar escores de segurança perceptual a partir de comparações pareadas de imagens. A segunda etapa diz respeito ao **classificador UrbanFormer**: um modelo que integra uma ConvNeXt Liu et al. (2022b) modificada (doravante CNN) com características semânticas do modelo de segmentação OneFormer Jain et al. (2023) para prever a segurança percebida. A terceira etapa descreve a técnica de **explicação do modelo**, que aplica o Grad-CAM exclusivamente ao ramo CNN em conjunto com as máscaras de segmentação para identificar e quantificar a contribuição de elementos visuais específicos às predições do modelo.

Diferentemente das abordagens que estimam a importância das características aplicando métodos *post-hoc*, como LIME ou SHAP, ao vetor de razão de pixels derivado da segmentação, este método de explicação opera diretamente no espaço da imagem: os mapas de atenção do Grad-CAM do ramo CNN são intersecados com as próprias máscaras de segmentação, de modo que a relevância é medida como sobreposição de área espacial, e não como um escore de importância sobre um vetor de características abstrato.

4.3.1 Quantificação da Percepção Urbana

Apoiamo-nos no conjunto de dados Place Pulse 2.0, apresentado no Capítulo 2. De suas seis categorias perceptuais, este capítulo foca exclusivamente na *segurança*, o atributo com o maior número de comparações pareadas (368.926). Os resultados pareados são convertidos em Q-scores em nível de imagem usando o algoritmo Strength of Schedule (SOS) (Park and Newman, 2005), definido no Capítulo 2 (Equações 2.19a–2.19d) e escalonado de 0 (muito inseguro) a 10 (muito seguro). Os Q-scores são computados globalmente em todas as imagens.

4.3.2 Localização de elementos urbanos usando segmentação semântica

Aplicamos a segmentação semântica para identificar elementos visuais-chave dentro das imagens, utilizando modelos avançados que classificam cada pixel em categorias específicas. Essa abordagem permite o isolamento e a análise precisos de elementos individuais, aprimorando nossa compreensão da distribuição espacial e das características detalhadas dos componentes visuais. Ao aproveitar a segmentação semântica, alcançamos uma percepção mais profunda e estruturada do conteúdo das imagens, apoiando uma análise mais abrangente dos dados visuais. Para essa tarefa, selecionamos os modelos de segmentação mais comumente usados nos estudos de percepção urbana: PSPNet (Zhao et al., 2017), DeeplabV3+ (Chen, 2017) e SegFormerB5 (Xie et al., 2021), identificados

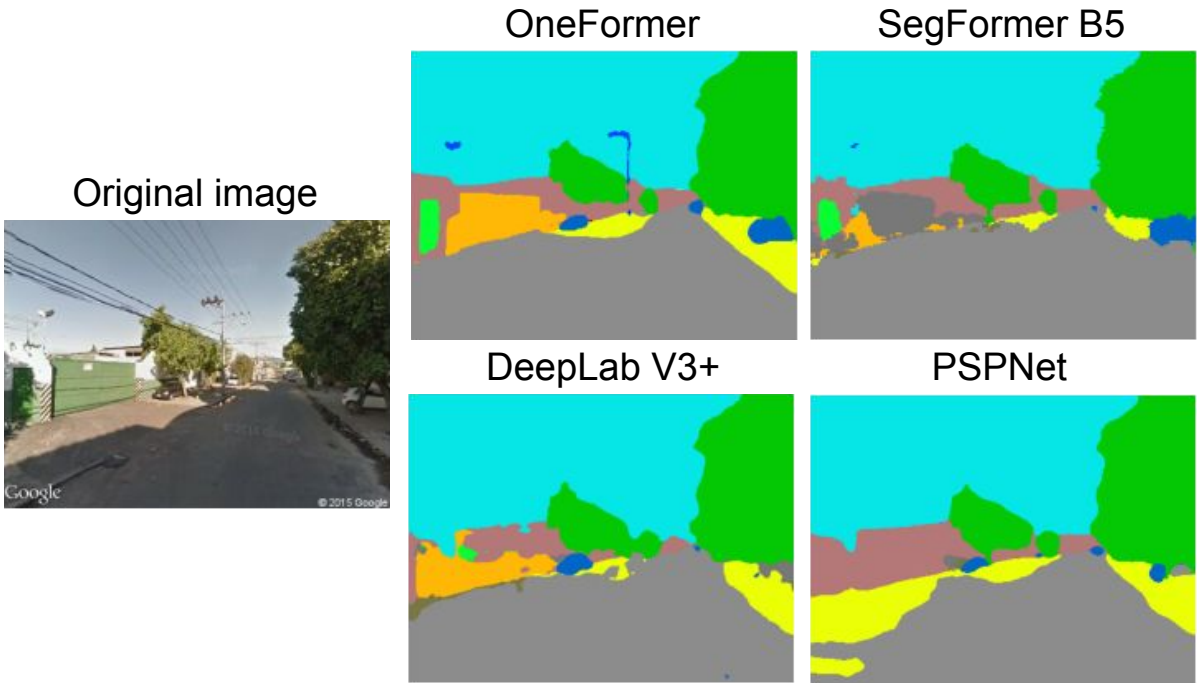

Figure 11 – Máscara de segmentação obtida a partir de OneFormer, SegFormer B5, PSPNet e DeepLabV3+. Observamos que alguns elementos visuais, como poste de luz, carros, portões e muros, são identificados incorretamente. Fonte: o autor.

no Capítulo 3. Também os comparamos com o modelo de segmentação OneFormer (Jain et al., 2023). Ao comparar o SegFormer B5, o OneFormer, o PSPNet e o DeepLab v3 para tarefas de segmentação semântica, observamos que o OneFormer lida com detalhes mais finos, especialmente em imagens de rua; outros modelos podem não capturar tão bem os contornos intrincados. Além disso, o OneFormer reportou melhores métricas na tarefa de segmentação do que os demais.

O OneFormer é projetado para lidar tanto com a segmentação semântica quanto com a de instâncias em um único modelo adaptável. Isso permite que o OneFormer capture relações espaciais sutis e detalhes mais finos com alta precisão, tornando-o a escolha preferida para aplicações de segmentação orientadas a detalhes. Para avaliar qualitativamente o desempenho da segmentação, selecionamos aleatoriamente 100 imagens segmentadas produzidas por cada método e inspecionamos manualmente as máscaras resultantes. Essa inspeção focou na correta delimitação dos contornos dos objetos e na segmentação acurada de elementos urbanos comuns, como calçadas, ruas, postes, veículos, muros e vegetação. A Figura 11 apresenta máscaras de segmentação representativas geradas pelos modelos avaliados. Nossa inspeção manual indica que o SegFormer, o PSPNet e o DeepLabV3+ frequentemente falham em segmentar postes de luz, mesclam incorretamente muros, calçadas e portões em uma única região e produzem segmentações de veículos incompletas ou imprecisas. Em contraste, o OneFormer fornece consistentemente segmentações mais detalhadas e acuradas, particularmente para objetos pequenos e contornos de cena complexos. Embora o OneFormer exija maiores recursos computacionais e envolva um pipeline de

treinamento e inferência mais complexo, sua qualidade de segmentação superior justifica esses custos adicionais para aplicações que requerem uma compreensão precisa da cena.

4.3.3 Classificador UrbanFormer

Propomos o UrbanFormer, um modelo de classificação que combina representações visuais e semânticas de cenas de *street view* para prever a segurança percebida. A arquitetura integra dois componentes pré-treinados: (i) um backbone CNN pré-treinado no ImageNet, que extrai mapas de características espaciais hierárquicos de imagens RGB brutas e produz um vetor de características compacto via global average pooling; e (ii) o modelo de segmentação OneFormer (Jain et al., 2023) pré-treinado no ADE20K, que produz um vetor de razão de pixels de tamanho 150, correspondente às 150 categorias semânticas definidas na taxonomia do ADE20K.

O componente CNN processa a imagem de entrada por meio de uma série de estágios convolucionais, cada um aumentando progressivamente a profundidade representacional enquanto reduz a resolução espacial, culminando em um mapa de características de alto nível que captura tanto padrões de textura locais quanto uma estrutura composicional mais ampla. O componente OneFormer produz a proporção de cada uma das 150 classes semânticas presentes na imagem, codificando sua composição semântica como um vetor de características de tamanho fixo. Essas duas representações são fundidas por concatenação e passadas a uma cabeça Perceptron Multicamadas (MLP) que realiza a classificação, treinada conjuntamente com uma perda de entropia cruzada. A Figura 10 ilustra a arquitetura do UrbanFormer.

A justificativa para essa fusão é que as características de aparência visual capturadas apenas pela CNN podem não capturar plenamente a composição semântica de uma cena — isto é, *quais objetos estão presentes* — que é independentemente informativa para a percepção de segurança. Ao codificar explicitamente a presença em nível de pixel de categorias semânticas como árvores, calçadas, cercas e edifícios, a contribuição do OneFormer permite que o modelo raciocine sobre o conteúdo da cena de uma maneira mais estruturada e interpretável.

4.3.4 Explicação do Modelo via Grad-CAM e IoU de Segmentação

Para identificar quais elementos visuais mais influenciam as predições de segurança, combinamos o método de explicação Grad-CAM com a saída de segmentação do OneFormer. O Grad-CAM é aplicado exclusivamente ao **ramo CNN** do UrbanFormer, uma decisão de projeto que decorre diretamente das propriedades estruturais de cada componente, como discutido a seguir.

Por que o Grad-CAM é aplicado apenas ao ramo CNN.

O Grad-CAM opera calculando o gradiente do escore da classe-alvo em relação aos mapas de ativação de uma camada convolucional escolhida e, então, ponderando esses mapas espaciais para produzir um mapa de calor de localização sobre a imagem de entrada. Sua aplicabilidade, portanto, exige que duas condições se sustentem simultaneamente: (1) a camada-alvo deve produzir mapas de características *espacialmente estruturados* de formato $H \times W \times C$, preservando a correspondência entre as posições das ativações e as regiões da imagem; e (2) o gradiente da saída de classificação deve fluir *contínua e ininterruptamente* pelo grafo computacional de volta até essa camada.

O backbone CNN satisfaz ambas as condições. Seus estágios convolucionais produzem mapas de características espaciais hierárquicos que mantêm a correspondência espacial com a imagem de entrada, e o gradiente do classificador MLP compartilhado propaga-se limpamente de volta, por meio do global average pooling, para esses mapas. Enganchar-se no último estágio convolucional da CNN, portanto, produz um mapa de calor bem definido e espacialmente significativo que reflete a decisão conjunta do modelo — moldada tanto pelo ramo visual quanto pelo semântico — para a classe de segurança-alvo.

O ramo OneFormer, por outro lado, não satisfaz essas condições no contexto do pipeline do UrbanFormer. Embora o OneFormer produza internamente representações espaciais ricas durante a segmentação — incluindo mapas de características do decodificador de pixels e embeddings de máscara — o que de fato é extraído e passado ao classificador MLP é o vetor de razão de pixels: um histograma de 150 dimensões que resume a proporção de cada classe semântica presente na imagem. Essa operação colapsa toda a estrutura espacial, transformando uma saída de segmentação bidimensional em um vetor plano e sem ordem. Nenhuma correspondência espacial entre as dimensões do vetor e as regiões da imagem é preservada após essa agregação.

Além disso, a implementação mais comum dessa extração de vetor semântico envolve uma operação de arg max sobre os mapas de probabilidade de classe para produzir uma máscara de segmentação rígida, seguida de contagem de pixels por classe. A função arg max é não diferenciável: ela não admite um gradiente bem definido e, portanto, cria uma descontinuidade no grafo computacional. Mesmo que as camadas convolucionais internas do OneFormer sejam parcialmente treináveis, os gradientes da perda de classificação binária não podem propagar-se pelo arg max de volta a essas camadas. Como resultado, aplicar o Grad-CAM às ativações internas do OneFormer não produziria mapas de calor que fossem significativamente atribuíveis à decisão de classificação; na melhor das hipóteses, tais mapas refletiriam a importância das características para o objetivo de *segmentação*, e não para a predição binária de segurança posterior.

Em resumo, o Grad-CAM é aplicável ao ramo CNN porque este produz mapas de características espaciais por meio de um pipeline diferenciável conectado de ponta a

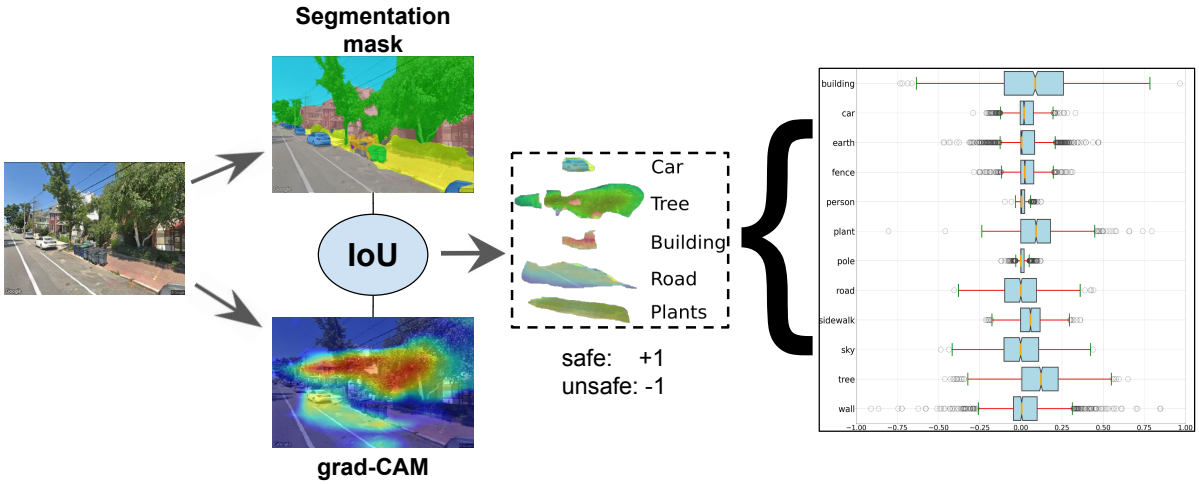

Figure 12 – Nossa técnica de explicação combina a máscara de segmentação e o método Grad-CAM para analisar a relação entre a percepção de segurança e a presença de elementos visuais. Fonte: o autor.

ponta ao classificador. O ramo OneFormer, por projeto, reduz sua saída a um descritor semântico não espacial por meio de uma operação não diferenciável antes de chegar ao classificador, tornando a atribuição espacial baseada em gradiente inaplicável a ele.

Formulação do Grad-CAM.

A técnica Grad-CAM, formalmente introduzida na Seção 2.2, calcula os pesos de importância dos neurônios para uma classe-alvo c e um mapa de características A^k como:

$$\alpha_k^c = \underbrace{\frac{1}{Z}\sum_i\sum_j}_{\text{GAP}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{grad-backprop}} \tag{4.1}$$

onde α_k^c representa o peso de importância do neurônio para o mapa de características k e a classe c (consistente com a notação introduzida no Capítulo 2); $Z = u \times v$ é o tamanho espacial do mapa de características; A_{ij}^k denota o valor de ativação na posição (i, j) no último estágio convolucional da CNN; e y^c é o escore previsto para a classe c produzido pelo classificador MLP. O mapa de localização resultante destaca as regiões da imagem, no espaço de características da CNN, mais relevantes para a predição do modelo para a classe de segurança-alvo.

É importante notar que, como o gradiente se origina do classificador MLP *fundido*, e não de uma cabeça *apenas de CNN*, o mapa de calor resultante é influenciado pela decisão conjunta de ambos os ramos. Em outras palavras, a atribuição espacial sobre os mapas da CNN codifica implicitamente como a composição semântica capturada pelo OneFormer contribuiu para moldar a atenção Grad-CAM do modelo sobre as regiões da imagem, ainda que o próprio ramo OneFormer não possa ser diretamente visualizado pelo Grad-CAM. A Figura 12 ilustra esse pipeline de explicação.

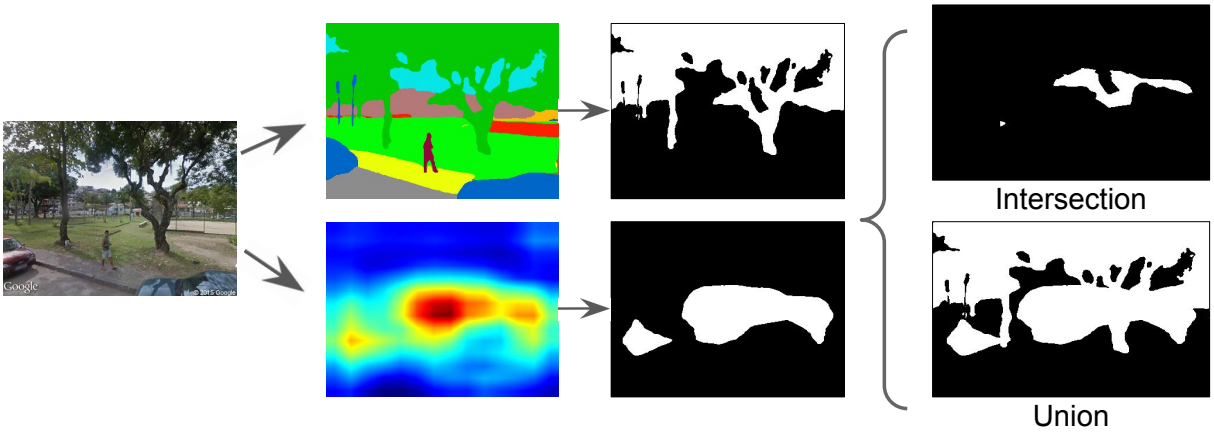

Figure 13 – Ilustração do cálculo da Interseção sobre União (IoU) para avaliar a qualidade da segmentação. O mapa de segmentação semântica é usado para extrair a máscara de referência (*ground-truth*) da classe-alvo (topo), enquanto o mapa de atenção/saliência é limiarizado para produzir a máscara prevista (base). O escore IoU é então calculado como a razão entre a interseção (pixels sobrepostos) e a união (pixels combinados) das duas máscaras binárias. Fonte: o autor.

IoU de segmentação para atribuição em nível de elemento.

Para quantificar a importância dos elementos semânticos individuais, calculamos a **Interseção sobre União (IoU)** entre cada máscara de segmentação produzida pelo OneFormer e o mapa de ativação do Grad-CAM gerado pelo ramo CNN. Um limiar de Grad-CAM de 0,3 é aplicado para reter apenas as regiões suficientemente ativadas, seguindo trabalhos anteriores (Vago et al., 2021; Li et al., 2025). O escore de IoU para cada categoria semântica indica quanto da área daquele objeto se sobrepõe às regiões a que o modelo atende ao fazer sua predição. Pesos positivos são atribuídos às predições de imagem segura e pesos negativos às predições inseguras, permitindo a análise de quais elementos estão associados a cada classe. Como mostrado na Figura 13, a métrica IoU é computada comparando a máscara binária derivada do mapa Grad-CAM com a máscara de segmentação de referência (ground-truth) correspondente, onde a sobreposição define a interseção e a área combinada define a união.

4.4 Experimentos

4.4.1 Exploração e Preparação do Conjunto de Dados

Durante a preparação do conjunto de dados, identificamos 2.471 localizações — definidas como posições que compartilham coordenadas idênticas de latitude e longitude — atribuídas a mais de um ID de imagem. Exemplos representativos incluem 130 localizações repetidas em Santiago (Chile), 126 em Berlim (Alemanha) e 112 em Montreal (Canadá), até

5 em Tel Aviv (Israel) e Seattle (EUA). Mapear todos os IDs duplicados para localizações únicas reduz o tamanho efetivo do conjunto de dados de 111.390 para **108.820 imagens**.

Também observamos um desequilíbrio significativo no número de imagens entre as cidades: algumas cidades, como Amsterdã, contribuem com apenas 622 imagens, enquanto Atlanta contribui com 3.965. Dado esse desequilíbrio, todos os experimentos são conduzidos usando o conjunto de dados multicidade completo.

Como etapa final de pré-processamento, retivemos apenas imagens com pelo menos cinco comparações pareadas. Esse critério de filtragem ajuda a reduzir a concentração de valores de escore específicos, como 3,333 e 6,666, que surgem de um número limitado de comparações e podem introduzir viés na distribuição de escores. Mais detalhes sobre essa questão podem ser encontrados em Moreno-Vera et al. (2021).

O conjunto de dados é particionado em 75% para treinamento e validação e 25% para teste, usando uma divisão estratificada por classe, de modo que a proporção segura/insegura seja preservada entre as partições. Para evitar o vazamento de dados, o colapso das posições de coordenadas duplicadas em IDs de imagem únicos descrito acima é realizado *antes* do particionamento, de modo que nenhuma localização física apareça tanto no conjunto de treinamento quanto no de teste; a divisão é definida sobre essas localizações sem duplicatas. Duas tarefas de classificação são definidas com base nos Q-scores derivados: (i) **classificação binária** (segura vs. insegura, dividida na pontuação mediana); e (ii) **classificação de 10 rótulos**, em que os intervalos de Q-score [0–1), [1–2), . . ., [9–10] são mapeados para os rótulos inteiros de 0 a 9. A seleção do modelo é realizada exclusivamente dentro da divisão de treinamento e validação por meio de validação cruzada estratificada de 5 dobras (5-fold) (Seção 4.4.2), deixando a partição de teste intocada durante o treinamento e a busca de hiperparâmetros.

4.4.2 Treinamento e Desempenho do Modelo

Todos os pesos do modelo são inicializados usando o critério uniforme de Xavier. A seleção do modelo é realizada por meio de busca em grade com validação cruzada estratificada de 5 dobras na divisão de treinamento e validação, para preservar as proporções de classe entre as dobras.

As Tabelas 6 e 7 reportam o desempenho de classificação do UrbanFormer e dos métodos anteriores nas tarefas de classificação de segurança binária e de 10 rótulos, respectivamente. Para a classificação binária, o UrbanFormer-CNN-only já alcança um desempenho competitivo, enquanto o UrbanFormer completo (CNN + OneFormer) atinge **75,22%** de acurácia, superando todas as baselines comparadas. Embora a classificação binária seja útil para distinguir cenas amplamente seguras e inseguras, ela comprime os escores de percepção originais em apenas duas categorias, descartando grande parte da informação subjacente.

Table 6 – Comparação de acurácia usando classificação binária (segura vs. insegura)

Model	Acc (%)
DSAPN+ResNet (Zhang et al., 2021)	64.87
MTDRALN-TC (Min et al., 2020)	65.82
VGG-GAP+Places365 (Moreno-Vera et al., 2021)	66.96
VGG19+ImageNet (Beaucamp et al., 2022)	67.01
PSPNet+SVR (Zhang et al., 2018a)	70.63
DeiT+ResNet50 (Sangers et al., 2022)	71.16
UrbanFormer-CNN-only (Ours)	**71.39**
UrbanFormer-CNN+OneFormer (Ours)	**75.22**

A tarefa de classificação de 10 rótulos fornece uma avaliação substancialmente mais fina ao particionar os escores contínuos de percepção de segurança em dez intervalos ordenados, preservando a distribuição dos escores e exigindo que o modelo discrimine entre diferenças sutis na segurança percebida. Sob esse cenário mais informativo, o UrbanFormer (CNN + OneFormer) alcança **78,51%** de acurácia, excedendo ou igualando todas as abordagens anteriores, incluindo modelos de fusão baseados em segmentação. O forte desempenho na tarefa de 10 rótulos indica que o UrbanFormer captura efetivamente não apenas distinções grosseiras de segurança, mas também as pistas visuais matizadas associadas a mudanças graduais na percepção urbana.

Table 7 – Comparação de acurácia usando classificação de 10 rótulos

Model	Acc (%)
SegFormerB5+RF (Zhao et al., 2024)	42.80
VGG19 (Zhao et al., 2024)	75.20
ResNet50 (Ma, 2023)	75.33
ConvNeXt-B (Zhao et al., 2024)	76.40
SFB5+ConvNeXt-B+RF (Zhao et al., 2024)	78.10
UrbanFormer-CNN-only (Ours)	72.19
UrbanFormer-CNN+OneFormer (Ours)	**78.51**

De modo geral, esses resultados demonstram que incorporar explicitamente a composição semântica da cena por meio do vetor de razão de pixels do OneFormer melhora consistentemente o desempenho em relação aos modelos baseados apenas em aparência. Além disso, as representações convolucionais hierárquicas da CNN fornecem um backbone visual robusto, enquanto a informação semântica complementar possibilita uma predição mais acurada ao longo de toda a distribuição dos escores de segurança urbana, e não apenas de seus extremos binários.

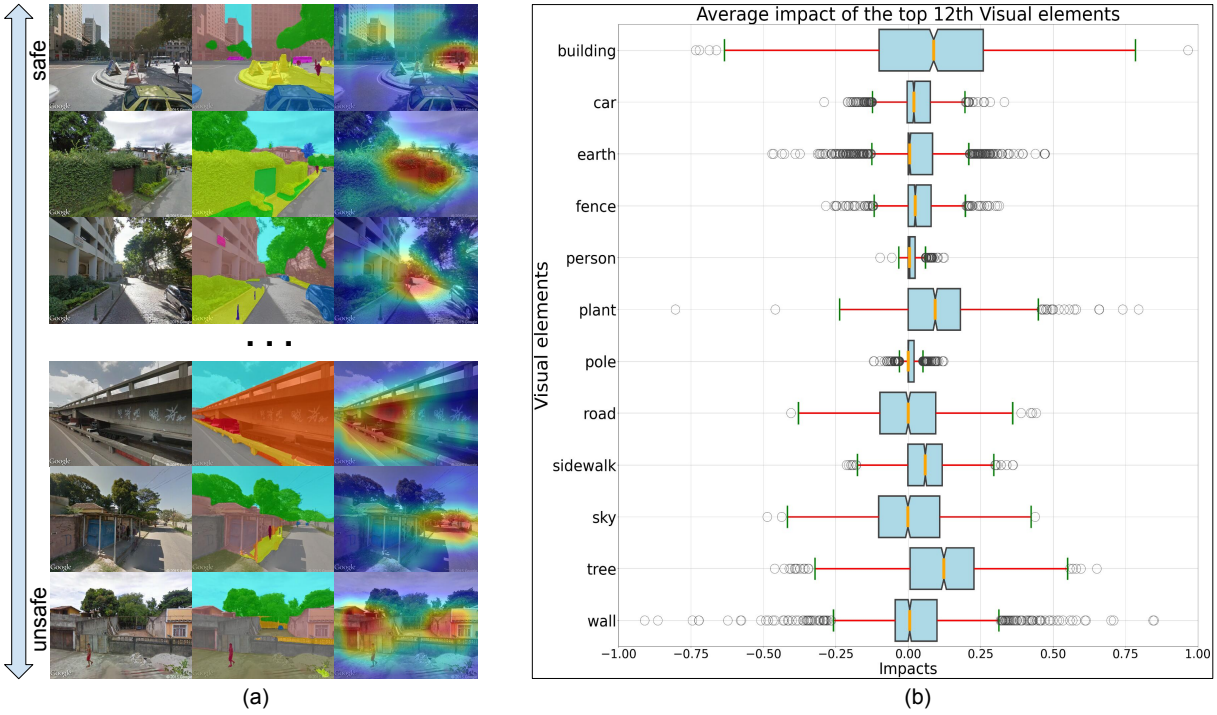

Figure 14 – Apresentamos as saídas do algoritmo Grad-CAM e a saída do OneFormer: (a) As 3 imagens mais seguras e as 3 menos seguras, a segmentação e as saídas do Grad-CAM. (b) O impacto médio dos 12 principais elementos visuais com presença em quase 90% das 108.820 imagens do conjunto de dados. Obtemos essa relevância realizando uma IoU entre as ativações e a máscara de segmentação, em que valores positivos correspondem a predições seguras e valores negativos correspondem a predições não seguras. Fonte: o autor.

4.4.3 Explicações do Modelo e Importância dos Elementos Visuais

Esta análise difere da prática mais comum de estimar a importância das características sobre um vetor de razão de pixels derivado da segmentação usando LIME ou SHAP; em vez disso, ela interseca a própria atenção Grad-CAM da rede com a máscara de segmentação, diretamente no espaço da imagem, para determinar quais elementos visuais o modelo de fato observou ao fazer cada predição.

A Figura 14(a) apresenta as três imagens mais seguras e as três menos seguras do conjunto de teste, ao lado de suas máscaras de segmentação do OneFormer correspondentes e das saídas do Grad-CAM computadas sobre o ramo CNN para o rótulo de referência de cada amostra.

Aplicando a análise de IoU entre as ativações do Grad-CAM do ramo CNN e as máscaras de segmentação em todas as 108.820 imagens, constatamos que os elementos visuais mais prevalentes — árvores, edifícios, ruas, calçadas, muros, cercas, solo e céu — estão presentes nas ativações de quase 96% das imagens do conjunto de dados. A Figura 14(b) reporta o impacto médio dos 12 elementos visuais mais prevalentes, com valores positivos correspondendo a predições seguras e valores negativos a inseguras.

A análise revela os seguintes achados-chave sobre como os elementos visuais se relacionam com a segurança percebida:

- **Elementos associados à percepção segura**: Árvores, grama, carros e plantas apresentam a mais forte sobreposição de IoU positiva em imagens seguras. Sua coocorrência consistente com a atenção do modelo em predições seguras sugere que a vegetação organizada e a infraestrutura conservada sinalizam segurança tanto para os observadores humanos quanto para o modelo.
- **Elementos associados à percepção insegura**: Muros, terra exposta (*dirt*), cercas e lixeiras exibem maior sobreposição de IoU em predições inseguras, sugerindo sua associação com a desordem e a negligência ambientais — um resultado que se alinha à teoria das Janelas Quebradas (Wilson and Kelling, 1982) discutida no Capítulo 2.
- **Elementos de alta prevalência e baixo impacto**: Elementos como ruas, edifício, calçadas e céu são altamente prevalentes em todas as imagens, mas exibem um impacto médio próximo de zero, indicando que sua *presença espacial* — a mera área que ocupam — é em grande parte não informativa para distinguir cenas seguras de inseguras. Ressaltamos que isso diz respeito à quantidade de céu visível, e não à *condição* do céu (clima, iluminação ou hora do dia), que as imagens estáticas de instantâneo único e a representação de razão de pixels usadas aqui não codificam, e que trabalhos anteriores mostraram moldar a percepção (Seção 4.4.4).
- **Elementos raros**: Das 150 categorias visuais identificadas pelo OneFormer, aproximadamente 120 aparecem em menos de 1% das imagens (por exemplo, flores, vasos, piscinas) e contribuem com um impacto médio próximo de zero para a análise.

Importante notar que a análise também sugere que a **ausência** de certos elementos (por exemplo, calçadas, vegetação) pode ser tão informativa quanto sua presença na caracterização de cenas inseguras. Elementos visuais com alta presença e um impacto médio de zero indicam relevância para *ambas* as categorias, segura e insegura, em igual medida; sua ausência, e não sua presença, pode ser a pista principal.

4.4.4 Limitações

Várias limitações deste estudo devem ser notadas. Primeiro, o conjunto de dados Place Pulse 2.0 não suporta análise em nível de cidade, uma vez que as comparações pareadas foram atribuídas aleatoriamente entre as cidades e o número de comparações por cidade é insuficiente para a modelagem específica por cidade. Segundo, o número de comparações por imagem é altamente desigual: a maioria das imagens foi comparada apenas três vezes, enquanto um pequeno número foi comparado até 78 vezes, introduzindo variabilidade na confiabilidade dos Q-scores individuais. Terceiro, o desequilíbrio de classes

presente após a atribuição de rótulos baseada em escores — tanto nas tarefas binária quanto de 10 rótulos — pode enviesar o desempenho do modelo em favor das classes sobre-representadas, mesmo após o aumento de dados.

Quarto, as restrições de recursos computacionais limitaram os tamanhos de lote e a extensão da busca de hiperparâmetros e de arquitetura, potencialmente afetando a capacidade de generalização do modelo. Quinto, embora o Grad-CAM forneça percepções qualitativas e quantitativas úteis sobre a quais regiões da imagem o modelo atende, ele reflete a atenção aprendida do modelo, e não um relato causal da percepção humana. Elementos que aparecem frequentemente em todas as imagens (por exemplo, céu, ruas) podem inflar seus escores de IoU independentemente de sua verdadeira relevância perceptual.

Uma ressalva relacionada diz respeito à interpretação de categorias de alta prevalência e baixo impacto, como o céu. Nossa análise mede apenas a *presença espacial* de cada elemento e trata o céu como uma categoria quase neutra, mas isso reflete uma propriedade da representação, e não da percepção em si. Como cada localização é capturada como um único instantâneo estático, a imagem mantém o clima, a iluminação e a hora do dia fixos, e o vetor de razão de pixels codifica quanto céu é visível, mas nada sobre sua aparência. A criminologia ambiental mostrou repetidamente que exatamente esses fatores importam: a segurança percebida e o medo do crime variam com o clima, a estação e a hora do dia, e os padrões de homicídio em São Paulo mostraram seguir variações espaço-temporais e climáticas (Ceccato, 2005, 2012; Ceccato and Loukaitou-Sideris, 2022). A neutralidade reportada aqui deve, portanto, ser lida de forma restrita — a *quantidade* de céu visível carrega pouco sinal discriminativo neste conjunto de dados — e não como evidência de que as condições do céu são perceptualmente irrelevantes. Capturar tais efeitos exigiria imagens que variam temporalmente e características do céu sensíveis à aparência, o que identificamos como uma direção importante para trabalhos futuros.

Sexto, a natureza baseada em presença da métrica de relevância IoU favorece sistematicamente elementos grandes e ubíquos em detrimento de elementos pequenos, porém salientes. Como a relevância é medida como sobreposição de área e, então, calculada em média ao longo do conjunto de dados, um elemento que ocupa apenas cerca de 1% de uma imagem (por exemplo, uma pichação ou uma placa danificada) contribui com um impacto médio próximo de zero, mesmo quando atrai fortemente a atenção do modelo sempre que aparece, enquanto elementos pervasivos, como árvores, acumulam alta sobreposição em grande parte em virtude de seu tamanho e frequência. A métrica, portanto, recompensa a prevalência em vez da saliência por objeto, e pistas pequenas de desordem podem ser subcreditadas em relação ao seu verdadeiro peso perceptual. Essa limitação, juntamente com possíveis correções, é discutida em maior profundidade como uma questão transversal no Capítulo 8.

Por fim, o escopo da explicação baseada em gradiente no UrbanFormer é inerente-

mente restrito ao ramo CNN, como discutido na Seção 4.3.4. O ramo OneFormer contribui para a decisão de classificação por meio de seu vetor de razão de pixels, mas essa contribuição não pode ser espacialmente visualizada via Grad-CAM devido à natureza não espacial e não diferenciável do processo de extração do vetor semântico. Trabalhos futuros devem explorar métodos de atribuição complementares — como Integrated Gradients ou SHAP — aplicados às dimensões do vetor semântico para desemaranhar as contribuições em nível de classe das categorias semânticas individuais daquelas capturadas pelos mapas de calor espaciais. Mecanismos que desemaranhem explicitamente a frequência de ocorrência da importância semântica também permanecem uma direção em aberto.

4.5 Considerações Finais

Este capítulo apresentou o UrbanFormer, um modelo fundido com características de segmentação semântica para a classificação da percepção de segurança urbana. A arquitetura combina as representações convolucionais hierárquicas da CNN com os vetores de razão de pixels produzidos pelo OneFormer, alcançando um desempenho competitivo tanto na tarefa de classificação binária quanto na de 10 rótulos no conjunto de dados Place Pulse 2.0, e superando abordagens anteriores baseadas em segmentação e em CNN.

Mais importante, a análise de explicação baseada em Grad-CAM e IoU — aplicada ao ramo CNN e cruzada com as máscaras de segmentação do OneFormer — fornece evidência interpretável de quais elementos visuais mais fortemente moldam a segurança percebida em SVI. Árvores, calçadas e vegetação organizada estão consistentemente associadas a percepções seguras, enquanto muros, solo exposto e lixeiras estão ligados a percepções inseguras. Esses achados fornecem uma base empiricamente fundamentada para recomendações de projeto urbano e vinculam o comportamento do modelo orientado a dados à literatura teórica mais ampla sobre desordem urbana.

Esses achados motivam o que se segue, mas saber que um elemento importa ainda não é saber o que uma rua precisaria para se sentir diferente: o Grad-CAM e a IoU mostram onde um modelo olha, e não como uma cena poderia plausivelmente mudar. O Capítulo 5 retoma exatamente essa questão, traduzindo vetores de diferença contrafactual — quais elementos visuais devem mudar — em recomendações em linguagem simples que podem apoiar o raciocínio em estágio inicial por parte dos planejadores urbanos. Para tanto, ele também introduz o UrbanPD4k, um conjunto de dados de 3.659 SVIs do Rio de Janeiro com anotações em nível de pixel de 13 elementos de desordem física não cobertos pelas taxonomias de segmentação padrão. O Capítulo 6 então vai ainda mais longe, apoiando-se tanto no vocabulário visual estabelecido aqui quanto no raciocínio contrafactual do Capítulo 5 para tornar tais mudanças visíveis e inspecionáveis, em vez de apenas descritas.

5 Explicações Contrafactuais e Interpretações Legíveis por Humanos da Percepção de Segurança Urbana

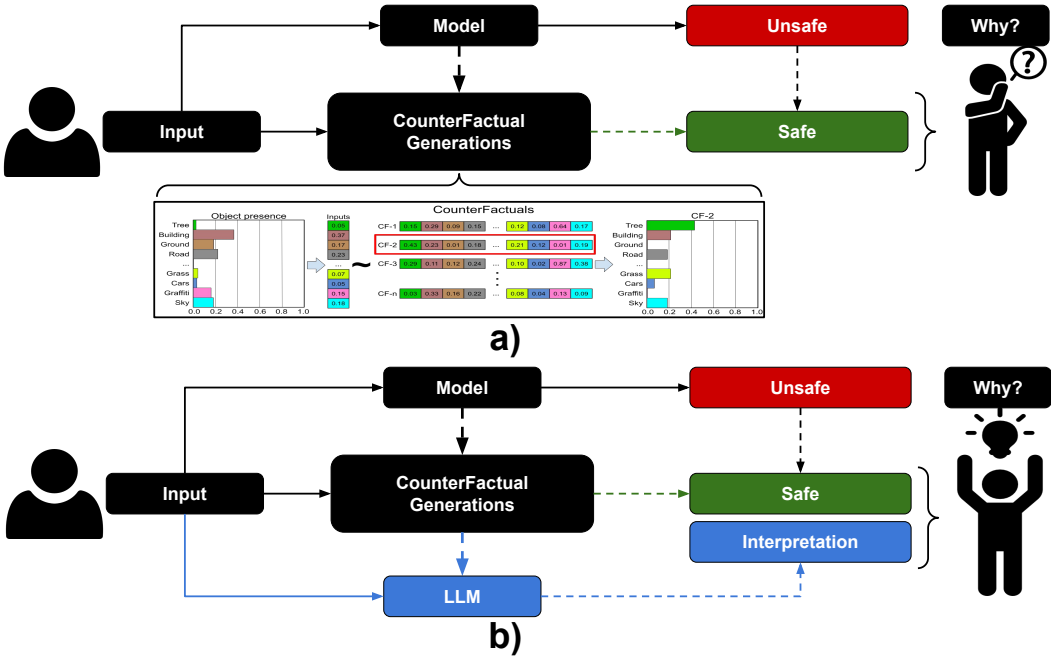

Figure 15 – Visão geral do UrbanCFX (*Urban CounterFactual eXplanations*) para predição e interpretação de segurança. O modelo inicialmente prevê um resultado inseguro. (a) Os contrafactuais sugerem mudanças na entrada que produzem uma predição mais segura, embora permaneçam difíceis de interpretar. (b) Um LLM gera explicações legíveis por humanos desses contrafactuais.

5.1 Introdução

O Capítulo 4 estabeleceu quais elementos visuais estão consistentemente associados à segurança percebida das ruas e quantificou sua relevância por meio da atenção espacial do modelo. Essa análise, no entanto, para no diagnóstico: saber que muros, terra exposta ou lixeiras importam não indica o que deveria mudar a respeito deles, nem produz recomendações sobre as quais planejadores ou formuladores de políticas possam agir diretamente. As explicações produzidas pelo Grad-CAM e pela IoU são numéricas e espaciais — escores de sobreposição atrelados a regiões da imagem — e interpretá-las ainda exige que um humano traduza mapas de calor e percentuais em intervenções concretas. Abordar essa limitação constitui a **segunda contribuição empírica** desta tese (Contribuição 3, introduzida no Capítulo 1). Especificamente, este capítulo identifica mudanças de cena que poderiam

plausivelmente melhorar a segurança percebida e as expressa como recomendações em linguagem natural que não especialistas podem prontamente compreender.

Duas limitações do vocabulário usado no Capítulo 4 motivam o primeiro passo deste capítulo. Taxonomias de segmentação padrão, como o ADE20K, descrevem o arranjo estrutural de uma rua — muros, calçadas, ruas — mas não distinguem um muro conservado de um coberto de pichação, nem uma calçada intacta de uma quebrada. No entanto, é precisamente essa distinção mais fina, bem documentada na literatura de percepção urbana desde a teoria "*Broken Windows*" (Wilson and Kelling, 1982), que mais diretamente sinaliza desordem a um observador humano (Nasar, 1998; Ulrich, 1979; Schroeder and Anderson, 1984). Este capítulo, portanto, estende o vocabulário semântico estabelecido anteriormente com um conjunto de elementos de desordem física anotados manualmente — pichação, muros danificados, cabos aéreos, calçadas quebradas, entre outros — capturando as pistas específicas mais associadas à percepção de negligência.

Com esse vocabulário mais rico estabelecido, o capítulo volta-se ao seu problema central: produzir uma explicação que uma pessoa, e não apenas um modelo, possa usar. A análise contrafactual identifica qual combinação mínima de mudanças de elementos deslocaria a classificação de uma cena de insegura para segura — mas a saída direta dessa análise é um vetor de diferença numérico, não mais imediatamente utilizável do que os mapas de calor do capítulo anterior. A principal contribuição metodológica deste capítulo fecha essa lacuna com um LLM: em vez de apresentar ao analista uma lista de características e valores com sinal, o pipeline aqui desenvolvido, o **UrbanCFX**, converte cada vetor de diferença contrafactual em uma sentença em linguagem simples que descreve o que mudou e por que isso importa. Este é o fio condutor que distingue este capítulo do anterior: o Capítulo 4 responde *quais* elementos importam e mostra essa resposta como um gráfico de atribuições; este capítulo responde *o que precisaria mudar* e expressa essa resposta como texto que uma pessoa pode ler diretamente, sem antes interpretar por conta própria um vetor, um escore ou um mapa de calor. Este trabalho foi publicado na *IEEE International Conference on Big Data* (BigData '25) (Moreno-Vera et al., 2025). Todos os materiais podem ser encontrados aqui[1].

Contribuições. Este capítulo contribui com: (i) o **UrbanPD4k**, um conjunto de dados de 3.659 SVIs do Rio de Janeiro com anotações manuais em nível de pixel de 13 elementos de desordem física não cobertos pelas taxonomias de segmentação padrão; (ii) uma comparação empírica mostrando que esses elementos de desordem carregam um sinal preditivo e contrafactual mais forte para a insegurança percebida do que as categorias gerais de segmentação; e (iii) o **UrbanCFX**, um pipeline que gera vetores de diferença contrafactual e, criticamente, os traduz em linguagem natural legível por humanos usando um LLM — transformando uma explicação numérica, voltada ao modelo,

[1] https://visualdslab.com/papers/UrbanPD4k/

em uma explicação baseada em texto, voltada ao humano. Esta última contribuição é a que mais diretamente faz avançar a tese como um todo: ela marca o primeiro ponto na progressão da dissertação em que um sinal derivado de máquina é expresso na mesma linguagem em que um planejador já raciocina, em vez de na linguagem do modelo que o produziu.

5.2 Trabalhos Relacionados

Há muito tempo, os elementos de desordem das ruas urbanas são reconhecidos como influentes na forma como as pessoas percebem e avaliam os ambientes urbanos. Sob a teoria "*Broken Windows*" (Wilson and Kelling, 1982), sinais visíveis de desordem — pichação, lixo, infraestrutura danificada — afetam negativamente as percepções de segurança e estabilidade social, e estudos em psicologia ambiental mostraram que os observadores interpretam tais pistas como sinais de negligência, moldando sentimentos de conforto e confiança (Nasar, 1998; Ulrich, 1979). Nasar e Jones (Nasar and Jones, 1997) constataram que bairros que exibem desordem física foram avaliados como significativamente inferiores em segurança e agradabilidade percebidas. Trabalhos computacionais mais recentes correlacionaram elementos de desordem específicos, particularmente a pichação, com indicadores socioeconômicos como o Índice de Desenvolvimento Humano (Diniz and Stafford, 2021; Alzate et al., 2021; Lavi. et al., 2022; Tokuda et al., 2019). Apesar dessa evidência, as categorias específicas de desordem permanecem em grande parte ausentes das taxonomias de segmentação usadas na modelagem convencional de percepção urbana, limitando sua disponibilidade para análises centradas na percepção — uma lacuna que o esforço de anotação deste capítulo aborda diretamente.

Do lado da explicação, o Capítulo 4 já discutiu como os modelos de percepção baseados em segmentação são comumente interpretados aplicando LIME ou SHAP a um vetor de características de razão de pixels ou baseado em presença (Zhang et al., 2018b; Min et al., 2020; Ma et al., 2025; Wu et al., 2025), produzindo um escore de importância por categoria semântica. Os métodos de explicação contrafactual vão um passo além, identificando o conjunto mínimo de mudanças de características que inverteria a predição de um modelo, em vez de meramente ordenar a importância das características (Mothilal et al., 2019), e foram aplicados na literatura mais ampla de IA explicável para gerar tais vetores de diferença para classificadores baseados em imagem (Gomez et al., 2020; Augustin et al., 2022). Em cada um desses casos, no entanto, a saída entregue ao analista permanece um vetor, uma tabela ou um mapa de calor — uma representação que ainda exige que o leitor traduza valores de características com sinal em um julgamento acionável. Separadamente, outros trabalhos usaram modelos de linguagem para descrever cenas urbanas diretamente, por meio de legendagem de imagens ou descrição multimodal (Ma, 2023; Ma and Wu, 2023), mas essas abordagens resumem *o que está presente* em uma

cena, em vez de *o que deveria mudar* para alterar como ela é percebida.

Diferenciação. O UrbanCFX ocupa o espaço que essas duas linhas de trabalho deixam em aberto. Ele não para em um vetor de diferença contrafactual, como fazem os métodos contrafactuais anteriores, nem usa um modelo de linguagem meramente para descrever uma cena, como fazem as abordagens baseadas em legendagem. Em vez disso, ele usa o modelo de linguagem como uma camada de tradução posicionada diretamente sobre a saída contrafactual: um *prompting* estruturado fundamenta o LLM no exato vetor de diferença produzido para uma dada imagem, e seu papel é especificamente renderizar esse vetor — numérico, estruturado e, de outra forma, legível apenas para alguém versado no espaço de características do modelo — em uma sentença em linguagem natural que descreve o que mudou e o que essa mudança significa para a segurança percebida. Onde o Capítulo 4 produziu uma explicação que permanece um gráfico, este capítulo produz uma que se lê como uma recomendação.

5.3 Metodologia

A metodologia está organizada em quatro estágios. O primeiro estágio envolve a **segmentação de imagem e anotação manual**: aplicar o OneFormer para extrair características semânticas padrão do ADE20K e enriquecê-las com elementos de desordem física anotados por humanos. O segundo estágio diz respeito à **seleção de conjunto de dados e de cidade**: motivar a escolha do Rio de Janeiro como cidade-alvo dentro do Place Pulse 2.0 e descrever o subconjunto de imagens usado em todos os experimentos. O terceiro estágio realiza a **explicação contrafactual**: gerar edições mínimas no espaço de características que deslocam a classificação de insegura para segura, sob duas representações de características complementares. O quarto e central estágio introduz a **interpretação legível por humanos orientada por LLM**: mapear vetores de diferença contrafactual para linguagem natural por meio de um *prompting* estruturado de um LLM.

5.3.1 Segmentação de Imagem e Anotações Manuais

Como no Capítulo 4, a segmentação semântica é realizada usando o OneFormer pré-treinado no ADE20K (Jain et al., 2023), produzindo um vetor de razão de pixels de 150 classes semânticas para cada imagem. Para este estudo, essas 150 classes são reagrupadas em oito categorias de **elemento visual** de nível mais alto — **Sky**, **Human**, **Construction**, **Floor**, **Vegetation**, **City elements**, **Terrain vehicles** e **Body of water** — para reduzir a dimensionalidade preservando a interpretabilidade semântica. A Tabela 8 resume as categorias reagrupadas.

Crucialmente, esse vocabulário padrão é estendido com anotações manuais em nível de pixel de **13 elementos de desordem física**, identificados por meio da literatura

Table 8 – Classes do ADE20K e novas categorias de grupo (elementos visuais)

Category	# classes	class name
Construction	14	wall, building, ceiling, house fence, column, skyscraper, bridge. bar, shack, tower, stadium fountain, outside door
Floor	7	floor, road, sidewalk ground, sand, path, land
Vegetation	6	tree, grass, plant, field flower, palm
Terrain vehicle	6	car, bus, truck, van motorcycle, bicycle
Body water	5	water, sea, river, waterfall, lake
City elements	4	signboard, streetlight pole, stoplight
Sky	1	sky
Human	1	person
Other	106	not included (less than 1% pixeles in images)

psicológica e sociológica sobre ambientes urbanos (Wilson and Kelling, 1982; Lynch, 1984; Schroeder and Anderson, 1984; Nasar, 1998). Esses elementos incluem **cabos aéreos, pichação, pavimento quebrado, sacos de lixo, latas de lixo, muros danificados, janelas quebradas, calçadas quebradas, placas de trânsito deterioradas, veículos policiais, indivíduos em situação de rua e indicadores relacionados de decadência física e vulnerabilidade social**.

Para esta etapa, dois anotadores treinados rotularam independentemente a presença, em nível de pixel, de cada pista em todas as n_i imagens, seguida de reconciliação para resolver desacordos (concordância entre avaliadores $\kappa = 0.81$). Ambos os anotadores são estudantes de doutorado — incluindo o autor desta tese — com experiência prévia de pesquisa em percepção urbana e familiaridade prática com o conjunto de dados Place Pulse, e ambos viveram no Rio de Janeiro por mais de dois anos. Essa formação compartilhada forneceu critérios consistentes e localmente fundamentados para identificar pistas de desordem específicas do contexto (como cabos aéreos densamente amontoados ou condições de muros de assentamentos informais) que um anotador não local poderia ignorar ou classificar erroneamente. Essas anotações permitem análises de granularidade fina sobre como a desordem visível se correlaciona com a segurança percebida, capturando qualidades de cena que nenhuma taxonomia de segmentação pré-existente cobre. A Figura 16 ilustra o processo de mesclar as máscaras de segmentação do ADE20K com as novas anotações de elementos de desordem.

A motivação para esse esforço de anotação decorre diretamente do Capítulo 4: elementos como terra exposta, lixeiras e muros estavam associados a percepções inseguras, mas os modelos de segmentação padrão não distinguem entre um muro bem conservado e

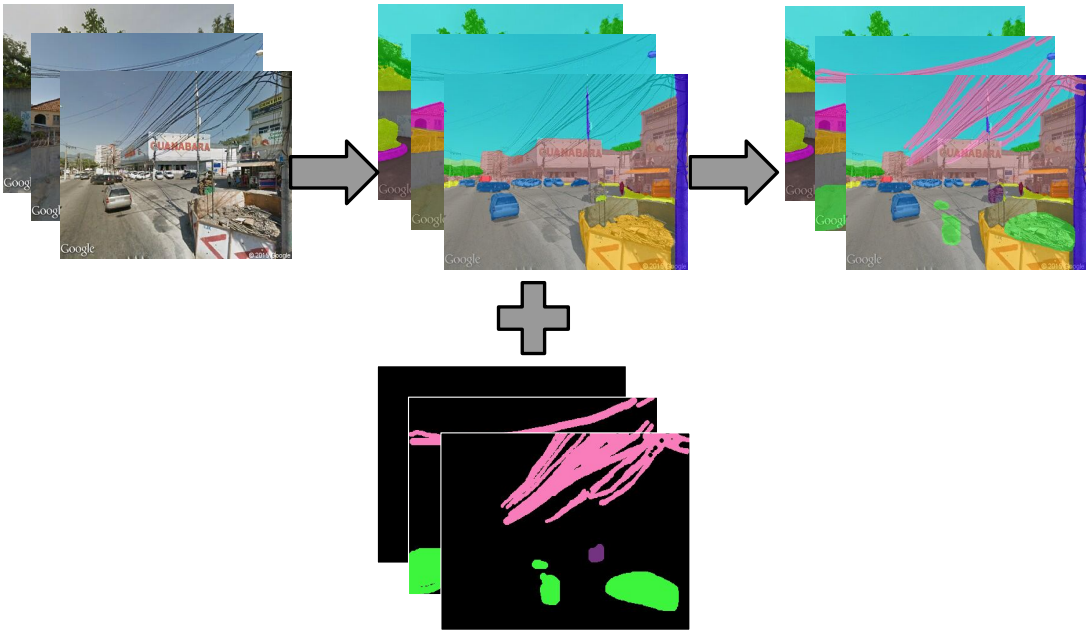

Figure 16 – Para gerar uma máscara de segmentação completa por amostra, realizamos uma substituição em nível de pixel. Fonte: o autor.

um coberto de pichação, ou entre uma calçada intacta e uma quebrada. As 13 anotações de desordem abordam precisamente essa lacuna representacional.

5.3.2 Seleção de Conjunto de Dados e de Cidade

O conjunto de dados usado neste estudo é extraído do Place Pulse 2.0, restrito a imagens da cidade do Rio de Janeiro, resultando em 3.659 cenas em nível de rua com seus resultados correspondentes de comparação pareada de segurança. A escolha do Rio de Janeiro não é arbitrária: entre todas as 56 cidades incluídas no Place Pulse 2.0, o Rio de Janeiro registra o menor escore médio de segurança percebida (Dubey et al., 2016; Moreno-Vera et al., 2021), tornando-o o contexto urbano mais crítico em termos de segurança percebida no conjunto de dados e, consequentemente, o mais informativo para estudar os elementos visuais associados a percepções inseguras. A Figura 17 ilustra a distribuição espacial da segurança percebida ao longo da cidade, juntamente com a cobertura geográfica das SVIs amostradas usadas neste estudo.

Esse perfil de baixa segurança percebida está fortemente ligado à alta prevalência de elementos de desordem física nas paisagens de rua do Rio. O tecido urbano da cidade é marcado por fortes contrastes socioespaciais (Diniz and Stafford, 2021): bairros abastados como Copacabana e Ipanema coexistem ao lado de extensos assentamentos informais (favelas), onde aproximadamente 23% da população vive em condições de infraestrutura deficiente e serviços públicos limitados (Liu et al., 2025). Essas áreas informais são caracterizadas por sinais visíveis de desordem física — cabos aéreos correndo entre os edifícios em densas teias, muros em deterioração, pichação, calçadas quebradas, acúmulo

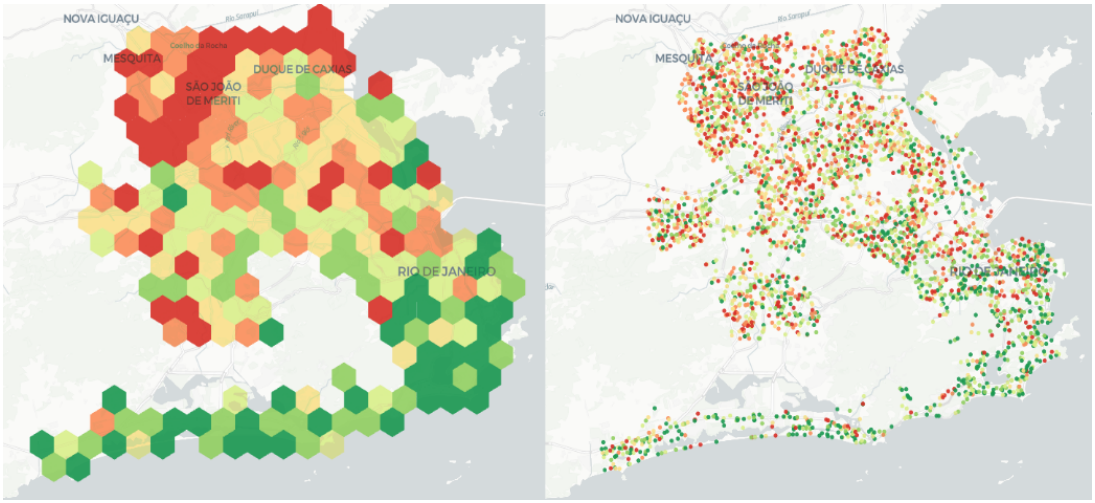

Figure 17 – À esquerda: Agregação hexagonal do escore mediano de segurança percebida (Q-score) calculado a partir de imagens de *street view* do Place Pulse 2.0, em que o verde indica maior segurança percebida e o vermelho indica menor segurança percebida. À direita: Distribuição geográfica das 3.659 imagens de *street view* amostradas, coloridas de acordo com seu rótulo binário de segurança (verde = seguro, vermelho = inseguro), ilustrando a cobertura espacial do conjunto de dados ao longo da cidade. Fonte: o autor.

de lixo e infraestrutura danificada — que são substancialmente mais prevalentes do que na cidade formal e que servem como as principais pistas visuais por meio das quais moradores e observadores formam percepções de segurança (Remigio et al., 2019). Essa concentração de elementos de desordem, precisamente aqueles capturados pelas 13 anotações manuais introduzidas neste capítulo, torna o Rio de Janeiro um contexto singularmente rico e representativo para estudar seu impacto na percepção de segurança.

De um ponto de vista metodológico, focar em uma única cidade também garante que todas as imagens compartilhem uma gramática visual coerente — a mesma mistura de estilos arquitetônicos, tipologias de infraestrutura e condições ambientais — evitando os confundidores intercidades inerentes ao conjunto de dados multicidade mais amplo analisado no Capítulo 4. Os escores de segurança perceptual para todas as imagens são computados usando o algoritmo Strength of Schedule (SOS), como descrito na Seção 2.4.2, e os rótulos de classificação binária são atribuídos com base no Q-score mediano.

5.3.3 Explicações Contrafactuais

As explicações contrafactuais, formalmente introduzidas no Capítulo 2 (Seção 2.2), buscam a mudança mínima em uma entrada que alteraria a predição do modelo. Aqui, os contrafactuais são gerados no **espaço de características**, em vez de no espaço de pixels: a entrada x é um vetor de descritores de elementos visuais, e o contrafactual x' especifica quais elementos devem mudar — em presença ou quantidade — para produzir uma classificação segura. Duas representações de características são avaliadas ao longo do

capítulo.

Presença/ausência binária.

Cada característica $x_i \in \{0, 1\}$ indica se um elemento visual está presente. O objetivo contrafactual minimiza o número de indicadores invertidos:

$$x' = \arg\min_{x' \in \mathcal{F}} \sum_{i=1}^{n} |x_i - x'_i| \quad \text{sujeito a} \quad f(x') \geq \tau \tag{5.1}$$

onde τ é o limiar de classificação da percepção de segurança e $\mathcal{F}$ é o conjunto de edições viáveis. A diferença $\Delta x_i = x'_i - x_i \in \{-1, 0, 1\}$ indica se o elemento i foi removido, mantido inalterado ou adicionado.

Razões de pixels.

Cada característica $x_i \in [0, 100]$ codifica a proporção da área da imagem ocupada por um elemento visual. O objetivo contrafactual busca o vetor mais próximo que alcança a classe-alvo:

$$x' = \arg\min_{x' \in \mathcal{F}} \mathcal{C}(x, x') + \lambda \cdot \mathcal{L}(f(x'), y') \tag{5.2}$$

onde λ equilibra a proximidade e o deslocamento da predição. A diferença de valor real $\Delta x_i = x'_i - x_i$ captura a magnitude da mudança de cobertura para cada elemento, permitindo um raciocínio contrafactual mais matizado do que as inversões binárias: a mudança de 10% para 25% de cobertura de árvores, por exemplo, é tratada como uma intervenção gradual, e não como uma simples alternância de presença.

5.3.4 Interpretações Legíveis por Humanos via LLM

A contribuição central deste capítulo é a transformação de vetores de diferença contrafactual em linguagem natural por meio de um LLM. Uma vez obtido o contrafactual x', as diferenças em nível de elemento $\Delta x = x' - x$ são estruturadas em dois conjuntos — elementos cuja presença ou cobertura aumentou e elementos cuja presença ou cobertura diminuiu — e passadas a um LLM por meio de um *prompt* estruturado. O LLM então produz uma recomendação em linguagem simples que descreve quais mudanças visuais melhorariam a segurança percebida da cena. O Prompt 1 ilustra o modelo usado para consultar o modelo, onde `added` e `removed` são preenchidos dinamicamente com as percepções contrafactuais.

Formalmente, esse mapeamento é expresso como:

$$\text{Interpretação} = \text{LLM}(\Delta x) \tag{5.3}$$

Para o caso binário, em que as diferenças são discretas, a explicação é derivada dos conjuntos de elementos adicionados e removidos:

$$\text{Explicação} = \text{LLM}(\{i \mid \Delta x_i = 1\},\ \{j \mid \Delta x_j = -1\}) \tag{5.4}$$

```
[System]: You are a person evaluating street images based on their visual appearance, tasked
    with improving street safety perception.

[User]: Imagine describing how a street scene might feel with visible changes in certain
    elements. We want you to describe which elements should be removed/reduced and which
    should be added/increased to improve the safety perception.

These changes are derived from comparing two representations of the same scene: original
    values and counterfactual values. The counterfactuals give us the following insights:

[Added or Increased Elements]: {\n.join(added) if added else 'None'}

[Removed or Decreased Elements]: {\n.join(removed) if removed else 'None'}

Based on these changes, please write a concise explanation (max 50 words) describing how these
     changes might affect your sense of safety.
```

Prompt 1 – *Prompt* de LLM utilizado para a avaliação da percepção de segurança das ruas

Para o caso de razão de pixels, um limiar ϵ filtra as flutuações negligenciáveis antes da interpretação:

$$\text{Explicação} = \text{LLM}(\{i \mid \Delta x_i > \epsilon\},\ \{j \mid \Delta x_j < -\epsilon\}) \tag{5.5}$$

garantindo que apenas mudanças semanticamente significativas cheguem ao modelo — por exemplo, a cobertura de pichação caindo de 15% para 2%, ou a cobertura de árvores subindo de 10% para 25% — em vez de ruído numérico.

Trabalhos anteriores compararam e avaliaram o desempenho de LLMs na geração de explicações textuais fundamentadas (Khan et al., 2024; Zytek et al., 2024), constatando que os modelos da família GPT-4 alcançam o desempenho mais forte. Com base nesses achados, usamos o GPT-4o mini, que fornece um equilíbrio favorável entre a qualidade da explicação e o custo computacional. Os prompts foram estruturados seguindo diretrizes estabelecidas para a geração de explicações em linguagem simples (Wang et al., 2024), instruindo o modelo a descrever as mudanças visuais em termos acessíveis aos planejadores urbanos, sem atribuir causalidade falsa nem introduzir informação ausente do vetor de diferença. Essa condição de fundamentação é essencial: o LLM não é solicitado a raciocinar livremente sobre a segurança urbana em geral, mas a traduzir uma intervenção específica, derivada do modelo, em linguagem natural.

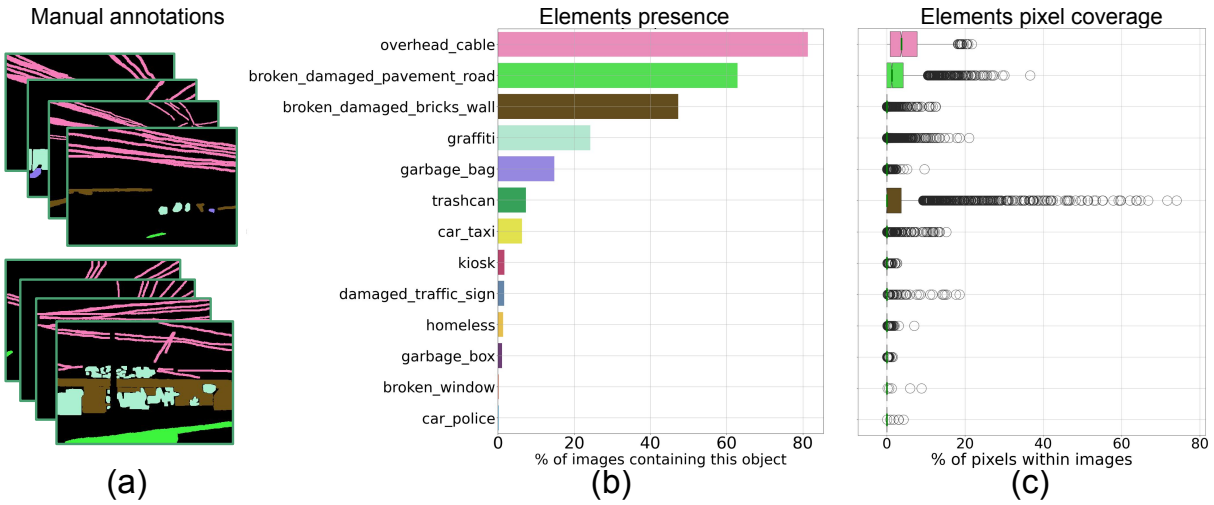

Figure 18 – Visão geral das estatísticas de anotação do conjunto de dados UrbanPD4k. (a) Amostras de máscaras de anotação em nível de pixel (b) Presença de elementos anotados, indicando a frequência de cada classe no conjunto de dados. (c) Cobertura de pixels dos elementos, mostrando a proporção da área da imagem ocupada por cada classe anotada. Em conjunto, essas visualizações destacam a diversidade e o equilíbrio do conteúdo anotado no UrbanPD4k. Fonte: o autor.

5.4 Experimentos

5.4.1 Conjunto de Dados UrbanPD4k e Estatísticas de Anotação

Introduzimos o conjunto de dados UrbanPD4k, que compreende 3.659 SVIs do Rio de Janeiro com anotações em nível de pixel em 13 categorias de desordem física. A Figura 18 resume as estatísticas de anotação por meio de três visualizações complementares. A subfigura (a) apresenta exemplos representativos das máscaras de segmentação geradas nas categorias anotadas, ilustrando a extensão do envolvimento humano no processo de rotulagem. A subfigura (b) reporta a frequência de ocorrência de cada classe anotada, fornecendo uma visão geral de quão frequentemente diferentes elementos de desordem física urbana — como cabos, latas de lixo, lixo e pichação — aparecem ao longo do conjunto de dados. A subfigura (c) apresenta a cobertura de pixels de cada classe, refletindo sua ocupação espacial relativa dentro das cenas. Em conjunto, essas visualizações mostram que grandes classes estruturais, como edifícios e ruas, respondem pela maior parte da área da imagem, enquanto elementos menores, porém semanticamente significativos — incluindo muros danificados, janelas quebradas, indivíduos em situação de rua e placas de trânsito deterioradas — ocorrem com menos frequência, mas enriquecem substancialmente a diversidade visual e contextual dos ambientes urbanos. Por fim, o conjunto de dados é dividido em 75% para treinamento e validação e 25% para teste em todos os experimentos, usando uma partição estratificada por classe que preserva a proporção segura/insegura e mantém o conjunto de teste totalmente reservado durante o ajuste do modelo, consistente com o protocolo seguro contra vazamento descrito no Capítulo 4.

5.4.2 Desempenho de Classificação e Impacto da Desordem

Quatro configurações de características são avaliadas tanto no cenário binário quanto no de razão de pixels: (i) classes do ADE20K isoladamente (150 características); (ii) ADE20K combinado com anotações de desordem (163 características); (iii) elementos visuais agrupados isoladamente (8 características); e (iv) elementos visuais agrupados combinados com anotações de desordem (21 características).

Um classificador Random Forest treinado sobre os vetores de características descritos acima serve como modelo de predição subjacente. Diferentemente da rede neural profunda usada no Capítulo 4, o Random Forest opera sobre características nomeadas e interpretáveis, tornando as diferenças contrafactuais diretamente mapeáveis a elementos visuais identificáveis e, portanto, adequadas para a interpretação baseada em linguagem. Também avaliamos modelos de classificação alternativos, incluindo métodos lineares, ensembles baseados em bagging e algoritmos de boosting. No entanto, essas abordagens alcançaram consistentemente um desempenho preditivo inferior em nosso conjunto de dados, tornando o Random Forest a escolha mais adequada para equilibrar a acurácia de classificação e a interpretabilidade.

Table 9 – Resultados de classificação usando configurações de valores binários e razões de pixels

Setting	Categories	Perception	Metrics		
			Precision	Recall	F-1
Binary values	ADE20K (150 classes)	not safe	0.59	0.65	0.62
		safe	0.60	0.55	0.57
	ADE20K+Disorder (163 classes)	not safe	0.64	0.65	0.65
		safe	0.67	0.60	0.63
	Visual elements (8 groups-classes)	not safe	0.52	0.42	0.46
		safe	0.51	0.61	0.55
	Visual elements+Disorder (21 classes)	not safe	0.66	0.71	0.67
		safe	0.65	0.66	0.66
Pixel ratios	ADE20K (150 classes)	not safe	0.67	0.73	0.69
		safe	0.71	0.63	0.67
	ADE20K+Disorder (163 classes)	not safe	0.70	0.76	0.73
		safe	0.73	0.68	0.70
	Visual elements (8 groups-classes)	not safe	0.69	0.72	0.70
		safe	0.70	0.64	0.66
	Visual elements+Disorder (21 classes)	not safe	0.72	0.77	**0.75**
		safe	0.75	0.69	**0.72**

A Tabela 9 reporta os resultados de classificação. Em ambas as representações de características, incorporar elementos de desordem melhora consistentemente o desempenho. No cenário binário, adicionar anotações de desordem ao vocabulário do ADE20K melhora os escores F1 para a detecção de insegurança em 2–5%; no cenário de razão de pixels, a melhoria alcança 5–7%. A melhor configuração geral — elementos visuais agrupados combinados com anotações de desordem no cenário de razão de pixels — alcança escores

F1 de 0,75 (não seguro) e 0,72 (seguro). Esses resultados confirmam que os elementos relacionados à desordem carregam um sinal preditivo significativo além do que as categorias gerais de segmentação fornecem, e que as representações de razão de pixels capturam uma variação perceptual de granularidade mais fina do que os indicadores binários de presença.

Para compreender melhor a contribuição das características de desordem, explicações SHAP (Lundberg and Lee, 2017) são aplicadas a ambas as representações. No cenário de razão de pixels, o SHAP atribui contribuições positivas e negativas a todos os elementos proporcionalmente à sua influência marginal: os elementos de desordem recebem consistentemente valores SHAP negativos, refletindo sua associação com uma menor segurança percebida, enquanto a vegetação e os elementos de construção bem conservados contribuem positivamente.

Criticamente, a magnitude dessas contribuições é assimétrica: os elementos de desordem impulsionam as predições inseguras mais fortemente do que sua ausência, ou a presença de elementos ordeiros, impulsiona as seguras — a desordem funciona principalmente como um sinal de perigo, em vez de sua ausência funcionar como um sinal igualmente forte de percepção de segurança. Isso é consistente com o enquadramento das Janelas Quebradas que motiva o esforço de anotação do UrbanPD4k: a negligência visível parece carregar mais peso perceptual do que a ordem visível.

No cenário binário, os elementos ausentes recebem uma atribuição SHAP próxima de zero, limitando a cobertura explicativa. **Essa assimetria reforça a representação de razão de pixels como a base mais forte tanto para a classificação quanto para a explicação**.

Essa análise baseada em SHAP, assim como as explicações baseadas em segmentação discutidas no Capítulo 4, permanece um relato agregado, em nível de característica: ela nos diz *quais elementos importam* ao longo do conjunto de dados, e não do que uma cena específica precisa. A subseção seguinte aborda diretamente essa lacuna em nível de instância.

5.4.3 Contrafactuais e Interpretações do LLM

A análise contrafactual identifica os elementos mais frequentemente modificados ao deslocar as predições de insegura para segura. Entre os elementos de desordem, os cabos aéreos, a pichação, os muros danificados, os sacos de lixo e as calçadas quebradas exibem a maior frequência de modificação nos 100 contrafactuais gerados por imagem usando a biblioteca DiCE (Mothilal et al., 2019). Entre os elementos gerais, a vegetação e a presença humana também estão comumente envolvidas. A frequência com que um elemento aparece nas transformações mínimas é interpretada como um indicador substituto de sua relevância contrafactual: os elementos nos quais o classificador mais se apoia para prever a insegurança são precisamente aqueles que mais frequentemente precisam mudar

para produzir uma predição segura.

Table 10 – Interpretações legíveis por humanos geradas por LLM para os contrafactuais

Image			
Δ safety probability	0.45	0.25	0.04
Objects Removed	graffiti, garbage, damaged sidewalk	overhead cables, damaged road	garbage bags
Objects Added	trees, walls, fences	Trees, grass	tree, grass, road
LLM interpretation	It's better to remove graffiti from walls, repair brick walls, and avoid overhead cables	The overhead cables should be removed, and adding grass and trees along the road will help increase people's sense of safety.	This street is already in good condition, but by incorporating more greenery and eliminating garbage bags, it could be improved even further.

A Tabela 10 apresenta interpretações geradas por LLM para três imagens inseguras representativas, correspondentes a lacunas de percepção alta, média e baixa (Δsafety). Para a imagem com o maior deslocamento (Δsafety = 0.45), em que pichação, lixo e uma calçada danificada são removidos e árvores, muros e cercas adicionados, o LLM produz: *"It's better to remove graffiti from walls, repair brick walls, and avoid overhead cables."* Para uma imagem moderadamente insegura (Δsafety = 0.25), em que cabos aéreos são removidos e grama e árvores adicionados, a interpretação diz: *"The overhead cables should be removed, and adding grass and trees along the road will help increase people's sense of safety."* Para uma imagem levemente insegura (Δsafety = 0.04), com apenas sacos de lixo removidos e vegetação e rua levemente adicionados, o modelo observa: *"This street is already in good condition, but by incorporating more greenery and eliminating garbage bags, it could be improved even further."*

As explicações calibram sua urgência à magnitude da lacuna de percepção: o caso de maior lacuna produz a intervenção mais prescritiva, enquanto o caso de menor lacuna reconhece a qualidade existente da cena. Esse comportamento emerge da estratégia de *prompting* fundamentado — o LLM responde ao vetor de diferença contrafactual real, e não a uma descrição geral da cena — e fornece evidência de que o pipeline produz recomendações contextualmente apropriadas, em vez de conselhos genéricos de projeto urbano.

5.5 Limitações

Várias limitações deste trabalho devem ser reconhecidas. Primeiro, o escopo geográfico é restrito ao Rio de Janeiro, e os achados sobre quais elementos de desordem mais importam podem não se generalizar diretamente para cidades com ambientes construídos ou normas socioculturais diferentes. Trabalhos futuros devem validar a abordagem em

múltiplas cidades para avaliar a robustez transcultural. Segundo, a anotação manual em nível de pixel de 13 elementos de desordem é demorada e sujeita à variabilidade do anotador. Escalar essa abordagem para coleções de imagens maiores exigiria estratégias de rotulagem semissupervisionada ou modelos automatizados de detecção de desordem.

Terceiro, a pichação é aqui anotada como uma única categoria de desordem, sem distinguir *tags* e vandalismo de muralismo e arte de rua encomendada. Isso é uma simplificação: nem toda pichação sinaliza negligência, e murais artísticos ou sancionados podem, em vez disso, ser percebidos como marcadores de cuidado, identidade ou vitalidade e lidos como *mais seguros*, e não como mais desordenados. Ao colapsar esses casos em uma única classe, o presente trabalho pode subestimar ou superestimar o peso perceptual da pichação dependendo do contexto da cena. Essa fusão, e suas implicações para o vocabulário de desordem, é examinada mais a fundo no Capítulo 8.

Quarto, os contrafactuais no UrbanCFX são gerados no espaço de características, e não no espaço de imagem: eles especificam mudanças nas razões ou nos indicadores de presença dos elementos visuais, mas não produzem imagens editadas. As intervenções propostas, portanto, não podem ser visualmente validadas quanto ao realismo ou à coerência perceptual por um observador humano. Essa limitação é diretamente abordada pelo UrbanCounterViz, introduzido no Capítulo 6, que gera imagens editadas fotorrealistas das mudanças propostas.

Quinto, embora a camada de interpretação do LLM produza explicações fluentes e fundamentadas, os modelos de linguagem podem ocasionalmente exagerar a causalidade ou introduzir inferências não estritamente justificadas pelo vetor de diferença. A estratégia de *prompting* estruturado mitiga esse risco, mas as interpretações resultantes devem ser entendidas como renderizações em linguagem natural de contrafactuais derivados do modelo, e não como relatos causais verificados.

Por fim, os escores de percepção de segurança subjacentes a todos os experimentos de classificação e contrafactuais são derivados de julgamentos pareados obtidos por *crowdsourcing*, que refletem a opinião pública agregada, e não as perspectivas dos moradores locais ou medidas de segurança percebida. Essa subjetividade introduz ruído e pode ser parcialmente abordada em trabalhos futuros integrando pesquisas locais, dados de criminalidade ou contexto demográfico.

5.6 Considerações Finais

Este capítulo apresentou o UrbanCFX, um pipeline que faz avançar a análise da percepção de segurança urbana em duas direções complementares. Empiricamente, o conjunto de dados UrbanPD4k e suas 13 anotações de desordem mostram que os elementos de desordem física — cabos aéreos, pichação, dano estrutural — carregam um sinal preditivo

e contrafactual mais forte para a insegurança percebida do que as categorias gerais de segmentação, e que as representações de razão de pixels fornecem uma fundamentação analítica mais rica do que os indicadores binários de presença. Metodologicamente, a camada de interpretação do LLM constitui a contribuição principal: ao converter vetores de diferença contrafactual em linguagem natural, ela fecha a lacuna comunicativa entre saídas interpretáveis por máquina e recomendações que podem apoiar o raciocínio em estágio inicial para planejadores urbanos e formuladores de políticas.

Ao lado do Capítulo 4, este capítulo estabelece duas perspectivas complementares sobre a análise interpretável de segurança urbana. O Capítulo 4 fundamenta a explicação na *atenção espacial do modelo* — quais regiões da imagem influenciam a predição — por meio do Grad-CAM e da IoU de segmentação. O presente capítulo fundamenta a explicação na *intervenção contrafactual* — quais mudanças de elementos deslocariam uma predição — por meio de contrafactuais, e externaliza essas intervenções como linguagem legível por humanos. Ambas as perspectivas, no entanto, operam sobre representações de características estáticas e não apoiam nem a exploração interativa de cenas nem a visualização fotorrealista das mudanças propostas.

Essas limitações remanescentes motivam diretamente a contribuição central de sistema desta tese. O UrbanCounterViz, introduzido no Capítulo 6, apoia-se na análise de elementos visuais e no vocabulário de desordem estabelecidos nestes dois capítulos, incorpora a edição de cenas fotorrealista orientada por VLM e incorpora todo o pipeline em um ambiente interativo de análise visual. Onde o UrbanCFX produz uma sentença de texto que descreve uma intervenção sugerida, o UrbanCounterViz renderiza uma SVI modificada e permite que o analista rastreie, compare e interrogue a transformação passo a passo. Ao fazê-lo, o UrbanCounterViz também adota e estende o vocabulário semântico desenvolvido no UrbanPD4k: as categorias de elementos de desordem identificadas aqui — cabos aéreos, pichação, muros danificados, calçadas quebradas — informam a análise de diferenças semânticas e a geração de ações de edição no cerne do pipeline contrafactual guiado por caminho do UrbanCounterViz, introduzido no Capítulo 6.

6 UrbanCounterViz: Um Sistema de Análise Visual para a Exploração Fundamentada da Segurança Urbana

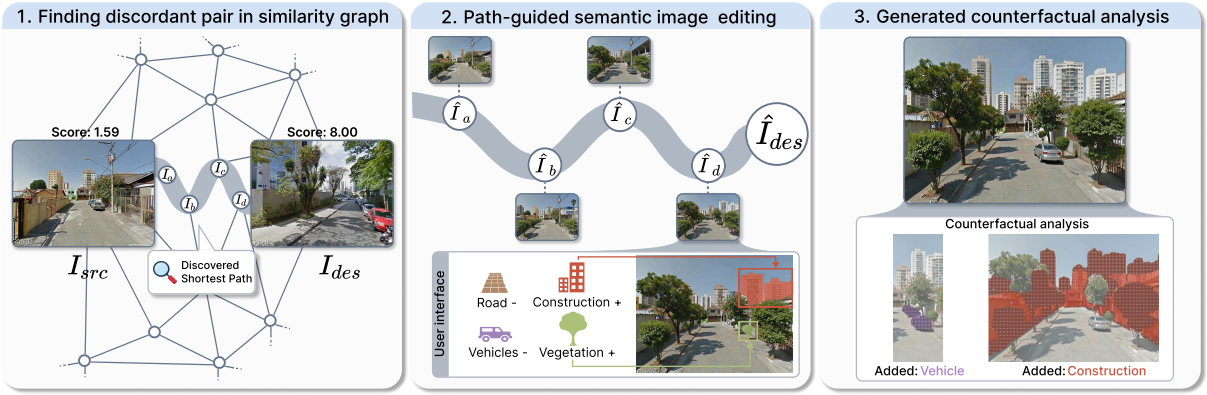

Figure 19 – Visão geral do UrbanCounterViz, resumindo as principais etapas do fluxo de trabalho de análise visual para a análise contrafactual de percepção de segurança. (1) Uma imagem de origem e uma imagem de destino com escores de percepção de segurança discordantes são selecionadas a partir do grafo de similaridade, e o caminho mais curto entre elas é calculado. (2) As diferenças semânticas ao longo desse caminho guiam a geração iterativa de imagens em direção a um objetivo de aumento ou diminuição da percepção de segurança. (3) Os contrafactuais gerados são então analisados por meio da composição da imagem, de indicadores urbanos e de visualizações interativas das edições introduzidas em cada etapa. Fonte: o autor.

6.1 Introdução

Os capítulos anteriores estabeleceram duas perspectivas complementares, porém incompletas, sobre a análise interpretável de segurança urbana. O Capítulo 4 demonstrou que elementos visuais específicos — árvores, calçadas, vegetação, muros, solo exposto e lixeiras — moldam consistentemente a atenção do modelo durante a predição de segurança, mas suas explicações são espaciais e estáticas: elas revelam *onde* um modelo olha, não *o que um analista deveria mudar*. O Capítulo 5 deu um passo em direção à acionabilidade ao gerar vetores de diferença contrafactual e convertê-los em recomendações em linguagem natural, mas suas intervenções existem apenas no espaço de características — elas especificam *o que* deveria mudar sem mostrar *como* essas mudanças ficariam em uma cena urbana real. Ambas as abordagens também compartilham uma limitação mais profunda: são pipelines não interativos que produzem uma única saída para uma dada entrada, sem nenhum mecanismo para que o analista explore transformações alternativas, inspecione etapas

intermediárias ou relacione mudanças visuais a resultados perceptuais em um ambiente unificado.

Este capítulo introduz o UrbanCounterViz, um sistema de análise visual projetado para fechar essas lacunas simultaneamente. Em vez de substituir o raciocínio em nível de característica estabelecido nos Capítulos 4 e 5, o UrbanCounterViz o estende para o domínio da imagem: as diferenças semânticas entre cenas urbanas reais são traduzidas em edições fotorrealistas por meio de um pipeline VLM de dois estágios, e todo o processo de transformação é tornado inspecionável por meio de uma interface interativa de análise visual. A Figura 19 apresenta uma visão geral da proposta.

A visualização e a análise visual são bem adequadas a esse desafio porque possibilitam a exploração integrada do conteúdo da imagem, da estrutura semântica, do comportamento do modelo e do conhecimento de domínio. Concomitantemente, as abordagens de edição de imagens contrafactuais e generativas oferecem uma direção promissora ao ilustrar como uma imagem poderia ser transformada para alterar o resultado de um modelo. No entanto, esses métodos frequentemente otimizam as edições com base nas respostas do modelo ou em geração aberta baseada em *prompts*, em vez de fundamentá-las em transformações observadas em cenas urbanas reais, e, consequentemente, podem produzir resultados visualmente impressionantes que são difíceis de interpretar como intervenções urbanas realistas. O UrbanCounterViz é projetado para ocupar exatamente essa lacuna: em vez de detectar elementos isoladamente ou gerar edições a partir de um prompt irrestrito, ele constrói *edições urbanas contrafactuais fundamentadas* a partir de *diferenças semanticamente significativas* identificadas entre imagens reais com diferentes escores de percepção, de modo que cada intervenção proposta seja sustentada por evidência já presente nos dados, e não pela imaginação irrestrita de um modelo.

Contribuições. Este capítulo contribui com a terceira e última contribuição empírica desta tese (Contribuição 4, conforme enumerada no Capítulo 1): (i) um pipeline de edição contrafactual guiada por caminho que modela as cenas urbanas como nós em um grafo de similaridade, identifica caminhos entre cenas visualmente similares com escores de percepção diferentes e deriva edições semanticamente orientadas a partir das diferenças em nível de objeto observadas ao longo desses caminhos; (ii) o próprio UrbanCounterViz, um sistema de análise visual que orquestra a geração contrafactual, o destaque visual das regiões editadas, os resumos semânticos e as visões coordenadas para apoiar a análise com humano no processo sobre como as mudanças de cena se relacionam com a segurança percebida; e (iii) um fluxo de trabalho reutilizável de análise visual interpretável, avaliado no Capítulo 7, que mostra como os contrafactuais resultantes apoiam o raciocínio sobre os fatores semânticos associados à segurança percebida e a identificação de estratégias plausíveis para o aprimoramento de cenas.

6.2 Trabalhos Relacionados

A contribuição deste capítulo situa-se na interseção de duas áreas de pesquisa ainda não cobertas nesta tese: a análise visual para análise de imagens urbanas e os métodos generativos para edição de cenas urbanas. Os Capítulos 4 e 5 já posicionaram este trabalho em relação aos modelos computacionais para percepção urbana — classificadores baseados em aparência e em segmentação, e métodos de explicação contrafactual, respectivamente — de modo que esta seção foca nas duas áreas específicas do próprio UrbanCounterViz.

6.2.1 Análise Visual para Análise e Percepção de Imagens Urbanas

A visualização e a análise visual desempenharam um papel importante em ajudar planejadores e analistas a interpretar imagens urbanas em larga escala. O *StreetVizor* (Shen et al., 2017) introduziu um sistema interativo de análise visual para explorar formas urbanas em escala humana a partir de SVI, apoiando a comparação entre áreas de interesse e múltiplas escalas espaciais. O *Urban Mosaic* (Miranda et al., 2020) estendeu essa direção ao possibilitar a recuperação, o agrupamento e a comparação sobre grandes coleções de *street view* espacial e temporalmente densas, ajudando os profissionais a analisar a mudança urbana e a comparar bairros. De forma mais ampla, o VIS4ML (Sacha et al., 2019) caracteriza o espaço de projeto para o aprendizado de máquina assistido por análise visual, e sistemas como o *ExplainExplore* (Collaris and van Wijk, 2020) ilustram como a comparação e as visões coordenadas apoiam a compreensão do modelo; levantamentos recentes enfatizam a importância de fluxos de trabalho integrados com humano no processo para o aprendizado profundo explicável (Choo and Liu, 2018; La Rosa et al., 2023).

Apesar dessas contribuições, os sistemas de análise visual anteriores para imagens urbanas focaram principalmente na exploração, na recuperação e na auditoria descritiva de dados de *street view* — apoiando os analistas em compreender *o que* está presente em uma cena ou *como* os elementos estão distribuídos por uma cidade, mas não em raciocinar sobre *o que poderia ser mudado* para produzir um resultado perceptual diferente. A pesquisa em análise visual sobre IA explicável, por sua vez, abordou em grande parte cenários de predição tabular ou genéricos, sem enfrentar os desafios específicos da edição semântica de cenas em contextos urbanos. O UrbanCounterViz conecta essas duas linhas de trabalho diretamente, adotando o paradigma de exploração interativa e multiescala estabelecido pelos sistemas de análise visual urbana, ao mesmo tempo em que o estende com uma camada de análise contrafactual que permite aos analistas inspecionar edições semânticas fundamentadas e rastrear as modificações de volta a regiões específicas da imagem.

6.2.2 Métodos Generativos para Edição de Cenas Urbanas

Métodos generativos — incluindo GANs, VAEs, modelos de difusão e abordagens de *inpainting* — foram aplicados à edição de cenas urbanas para tarefas como embelezamento visual, modificação baseada em preferências e simulação de intervenções de renovação urbana. O *FaceLift* (Joglekar et al., 2020) demonstrou que as GANs poderiam gerar melhorias de fachada que aumentam a atratividade percebida. O UCA (Law et al., 2023) usou modelos generativos para explicar mudanças na atratividade urbana. Trabalhos mais recentes baseados em difusão incluem o UPD-Explainer (Hu et al., 2023), que gera imagens contrafactuais destacando sinais de desordem física; o VPI-Toolkit (Tan et al., 2025) e o URSimulator (Hu et al., 2025), que simulam intervenções de percepção visual. Os VLMs separadamente tornaram-se centrais para a computação urbana de forma mais ampla, usados para comparar julgamentos de máquina e humanos, inferir indicadores urbanos a partir de imagens e aprender representações multimodais para tarefas urbanas posteriores (Levering et al., 2024; Beneduce et al., 2025; Zhou et al., 2025b; Yao et al., 2026; Zhanga et al., 2024; Huang et al., 2024; Malekzadeh et al., 2024; Moreno-Vera and Poco, 2025).

Apesar desse progresso, as abordagens generativas para a edição de cenas urbanas compartilham limitações persistentes. Elas frequentemente fornecem controle limitado sobre as mudanças específicas introduzidas — as edições podem incluir transformações arquitetônicas irrealistas, inconsistências de iluminação ou deslocamentos de ponto de vista que comprometem a fidelidade da cena — e, mais criticamente, oferecem suporte mínimo para compreender *por que* uma mudança gerada é significativa ou *como* ela se relaciona com o objetivo perceptual pretendido, uma vez que as edições são otimizadas para alterar a saída de um modelo ou satisfazer um critério de estilo, em vez de serem fundamentadas em transformações observadas em ambientes urbanos reais.

Diferenciação. O UrbanCounterViz aborda as limitações de ambas as linhas de trabalho simultaneamente. Diferentemente dos sistemas de análise visual urbana anteriores, ele não para na exploração e na auditoria descritiva: ele adiciona uma camada contrafactual que permite aos analistas inspecionar edições semânticas fundamentadas e relacioná-las à segurança percebida. Diferentemente dos métodos de edição generativa anteriores, ele não sintetiza modificações hipotéticas nem otimiza edições para deslocar a saída de um modelo: cada intervenção é derivada de diferenças semânticas de fato observadas entre cenas urbanas reais e visualmente similares, e a edição resultante — juntamente com o grafo de similaridade, o vetor de diferença semântica e as ações de edição que a produziram — é exposta ao analista por meio de uma interface interativa e coordenada, em vez de entregue como uma imagem final opaca. Essa combinação de fundamentação, geração e inspeção interativa dentro de um único ambiente é o que distingue o UrbanCounterViz de todos os sistemas revisados acima.

6.3 Visão Geral do Sistema

Compreender como as variações nos elementos visuais influenciam as percepções humanas de segurança exige mais do que um modelo preditivo — exige um arcabouço para explorar interativamente essas relações, conectá-las a intervenções realistas e avaliar sua plausibilidade. Apoiando-nos no pipeline analítico estabelecido nos capítulos anteriores e em princípios consagrados de análise visual, como "*overview first, zoom and filter, then details-on-demand*" (Shneiderman, 2002), derivamos um conjunto de requisitos de projeto e tarefas analíticas que, coletivamente, enquadram o desenvolvimento do UrbanCounterViz.

Esses requisitos não foram definidos isoladamente. Eles foram informados pelas análises empíricas reportadas nos Capítulos 4 e 5, que estabeleceram o vocabulário de elementos visuais, o arcabouço de anotação de desordem e a estratégia de raciocínio contrafactual que o UrbanCounterViz estende para o domínio da imagem (Moreno-Vera et al., 2024, 2025). No entanto, esses estudos também expuseram uma limitação consistente: raciocinar exclusivamente por meio de representações agregadas ou tabulares torna difícil compreender como tais mudanças se manifestariam visualmente em uma cena real. Essa lacuna — entre a descrição em nível de característica de uma intervenção e sua aparência visual — motivou a transição para a análise baseada em imagem e o projeto do UrbanCounterViz como um sistema que conecta diferenças semânticas, edições de cena plausíveis e resultados perceptuais em um ambiente visualmente interpretável e interativo.

6.3.1 Requisitos de Projeto

Com base nas lacunas de pesquisa identificadas no Capítulo 3 e nas lições aprendidas com as contribuições empíricas dos Capítulos 4 e 5, estabelecemos quatro requisitos centrais (R) para guiar o projeto do UrbanCounterViz.

R1 — Contextualizar a Similaridade e as Transições entre Cenas. Os usuários devem ser capazes de selecionar SVI geolocalizadas, comparar cenas urbanas visualmente similares e inspecionar sequências de transição que as conectam. O sistema deve revelar como os atributos perceptuais e as categorias semânticas evoluem entre cenas similares e ao longo do caminho entre duas imagens selecionadas. Esse requisito responde diretamente à ausência de exploração interativa nos sistemas de percepção urbana anteriores (Shen et al., 2017; Miranda et al., 2020), que apoiam a auditoria descritiva de coleções de cenas, mas não o raciocínio sobre como as cenas transitam de uma para outra.

R2 — Direcionar Edições Semânticas. Os usuários devem ser capazes de inspecionar as diferenças semânticas entre imagens selecionadas e usar essas diferenças para guiar a geração contrafactual. O sistema deve fornecer edições semânticas orientadas ao planejamento e apoiar a geração de imagens contrafactuais guiadas. Isso estende o raciocínio contrafactual introduzido no Capítulo 5 do espaço de características para o domínio da imagem, onde as

intervenções podem ser avaliadas visualmente.

R3 — Inspecionar e Compreender Modificações Visuais. Quando uma imagem contrafactual é gerada, os usuários devem ser capazes de identificar o que mudou, onde mudou e como essas mudanças se relacionam com o deslocamento perceptual pretendido. O sistema deve apoiar a comparação visual detalhada, distinguindo claramente as categorias semânticas adicionadas, removidas e preservadas. Esse requisito aborda a opacidade dos métodos generativos anteriores (Joglekar et al., 2020; Hu et al., 2023, 2025), que produzem imagens editadas sem revelar a lógica da transformação.

R4 — Avaliar o Impacto Multidimensional. Os usuários devem ser capazes de avaliar os efeitos das edições em múltiplas dimensões simultaneamente. O sistema deve fornecer resumos visuais que comparem os atributos perceptuais (por exemplo, escores de percepção de segurança), os indicadores urbanos (por exemplo, openness, greenery, walkability) e as distribuições de categorias semânticas antes e depois da edição. Esse requisito conecta-se à necessidade mais ampla de ferramentas que apoiem recomendações de projeto urbano transparentes e fundamentadas em evidências, conforme identificado nas considerações finais do Capítulo 3.

6.3.2 Tarefas Analíticas

Apoiando-nos nesses requisitos, definimos um conjunto conciso de tarefas analíticas que o UrbanCounterViz é projetado para apoiar. Essas tarefas seguem um padrão de visão geral para detalhe (Shneiderman, 2002): os analistas inicialmente obtêm uma ampla visão geral espacial da coleção de imagens, depois examinam as similaridades e os escores perceptuais entre as cenas selecionadas e, por fim, inspecionam os detalhes do processo de edição e seus resultados.

T1: Explorar a similaridade e as transições entre cenas (R1).

- **T1.1** Identificar e comparar cenas geolocalizadas visualmente similares com base nos atributos perceptuais e na composição semântica.
- **T1.2** Analisar sequências de transição entre duas imagens selecionadas para compreender como as categorias semânticas e os escores perceptuais mudam ao longo do caminho.

T2: Analisar e direcionar edições semânticas por meio da geração contrafactual (R2).

- **T2.1** Identificar e quantificar as diferenças semânticas entre um par de imagens selecionadas e suas contrapartes geradas.

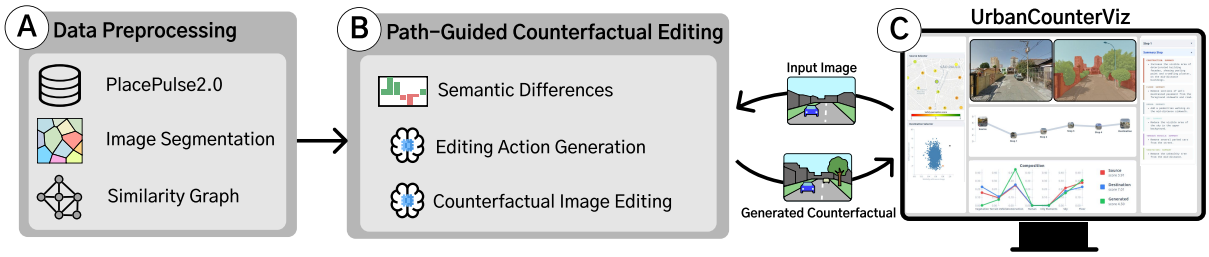

Figure 20 – Visão geral do fluxo de trabalho do UrbanCounterViz. (A) **Pré-processamento de dados**: transformamos as comparações pareadas do Place Pulse 2.0 em representações estruturadas, incluindo escores de percepção, categorias semânticas, atributos visuais profundos e similaridades entre imagens. (B) **Edição contrafactual guiada por caminho**: as diferenças semânticas entre pares de imagens são usadas para derivar ações de edição e gerar imagens contrafactuais fundamentadas. (C) **UrbanCounterViz**: um sistema interativo de análise visual que apoia a exploração de cenas urbanas, a inspeção de mudanças semânticas, a comparação de edições geradas e a avaliação de como essas edições se relacionam com a segurança percebida. Fonte: o autor.

- **T2.2** Inspecionar as etapas intermediárias de transição no processo de geração para compreender como as categorias semânticas e os atributos perceptuais mudam progressivamente e onde surgem as diferenças-chave.

T3: Comparar, interpretar e relacionar impactos multidimensionais (R3, R4).

- **T3.1** Identificar e caracterizar as modificações visuais — categorias semânticas adicionadas, removidas e preservadas — e relacioná-las às mudanças perceptuais.
- **T3.2** Relacionar indicadores urbanos, categorias semânticas, escores perceptuais e padrões de similaridade entre as imagens originais e as contrafactuais geradas.

Essas tarefas informaram diretamente o projeto das seis visões coordenadas do UrbanCounterViz, cada uma das quais atende a requisitos específicos, permanecendo integrada às demais para possibilitar uma exploração, comparação e interação fluidas.

6.3.3 Fluxo de Trabalho

A Figura 20 fornece uma visão geral de alto nível do fluxo de trabalho do UrbanCounterViz, que consiste em três estágios executados em sequência.

No **estágio de pré-processamento de dados** (Figura 20.A), as SVI brutas e os julgamentos humanos pareados são transformados em representações estruturadas: escores de percepção, embeddings de características visuais profundas, proporções de categorias semânticas, indicadores urbanos e um grafo de similaridade pareada entre imagens. Esse estágio é executado offline e produz as estruturas de dados que apoiam toda a interação subsequente.

No **estágio de edição contrafactual guiada por caminho** (Figura 20.B), o sistema calcula as diferenças semânticas ao longo dos caminhos do grafo entre uma imagem de origem selecionada pelo usuário e uma imagem de destino com um escore de percepção de segurança mais alto. Essas diferenças são usadas para derivar ações de edição explícitas, que são então passadas a um pipeline VLM de dois estágios para gerar imagens modificadas fotorrealistas. Crucialmente, esse estágio decompõe uma lacuna perceptual potencialmente grande em etapas menores e interpretáveis, cada uma fundamentada em diferenças observadas entre cenas adjacentes reais no grafo.

Por fim, a **interface de análise visual** (UrbanCounterViz, Figura 19.C) integra esses componentes computacionais em um ambiente interativo onde os usuários podem guiar a geração de imagens, inspecionar as mudanças semânticas e visuais em cada etapa e avaliar se os contrafactuais resultantes apoiam seus objetivos analíticos. A interface é a contribuição principal do sistema: ela torna o pipeline transparente, direcionável e interpretável, em vez de uma transformação de caixa-preta da entrada para a saída.

6.4 Pré-processamento de Dados

O estágio de pré-processamento converte os julgamentos humanos pareados brutos e as SVI nas representações estruturadas necessárias para a exploração interativa e a geração contrafactual. Ele compreende três etapas: (i) estimar os escores perceptuais em nível de imagem a partir das comparações pareadas, (ii) extrair características semânticas e computar indicadores urbanos a partir das imagens e (iii) construir um grafo de similaridade sobre a coleção de imagens.

6.4.1 Estimativa dos Escores de Percepção

Os resultados das comparações pareadas são convertidos em escores perceptuais em nível de imagem usando o algoritmo de ranqueamento Strength of Schedule (SOS), formalmente introduzido na Seção 2.4.2. Essa abordagem produz Q-scores, escalonados de 0 (muito inseguro) a 10 (muito seguro), que servem como fundamento perceptual para toda a seleção de cenas, o cálculo de caminhos e a avaliação contrafactual subsequentes dentro do UrbanCounterViz. Embora os Q-scores sejam computados globalmente em todas as imagens, a análise subsequente e a construção do grafo são realizadas independentemente para cada cidade.

6.4.2 Extração de Características Semânticas e Indicadores Urbanos

Para caracterizar o conteúdo visual de cada imagem, as proporções de categorias semânticas são extraídas usando o OneFormer pré-treinado no ADE20K (Jain et al., 2023) — o mesmo modelo de segmentação estabelecido nos Capítulos 4 e 5. Para cada imagem I,

a proporção de pixels $p_k(I)$ atribuída a cada uma das 150 classes semânticas do ADE20K é computada. Essas proporções fornecem um resumo conciso e comparável da composição da cena e são subsequentemente usadas para identificar diferenças semânticas entre imagens e para derivar ações de edição candidatas.

Como 150 categorias individuais são finas demais para a análise e inspeção interativas, elas são agregadas em nove grupos de nível mais alto: **Sky**, **Human**, **Construction**, **Floor**, **Vegetation**, **City elements**, **Terrain Vehicles** e **Other**. Esse agrupamento preserva as distinções semânticas mais relevantes para a percepção de segurança — conforme estabelecido no Capítulo 5 — ao mesmo tempo em que as apresenta em um nível de abstração que os analistas podem prontamente interpretar e sobre o qual podem agir. A categoria **Other** é excluída da análise interativa devido à sua representação mínima e heterogênea no conjunto de dados.

Além das proporções semânticas, seis indicadores de percepção urbana de nível mais alto são computados para cada imagem: **greenness**, **walkability**, **driving safety**, **imageability**, **enclosure** e **openness** (Ma et al., 2021; Qiu et al., 2023). Esses indicadores fornecem uma camada complementar de análise que vai além da composição em nível de objeto para capturar qualidades ambientais comumente associadas às percepções humanas dos espaços urbanos. Eles são exibidos ao lado das proporções semânticas na interface para apoiar a avaliação multidimensional das edições geradas (R4).

6.4.3 Pares Discordantes na Percepção Urbana

Um desafio recorrente na pesquisa em percepção urbana é compreender *por que* duas cenas visualmente similares recebem escores de percepção humana substancialmente diferentes. Os pipelines de predição padrão produzem um escore por imagem, mas não oferecem nenhum mecanismo direto para identificar quais diferenças visuais são responsáveis por julgamentos humanos divergentes. Para abordar isso, introduzimos o conceito de **pares discordantes** como uma contribuição formal desta tese. Embora a intuição por trás de comparar cenas similares com escores diferentes tenha motivado a análise de elementos visuais no Capítulo 4 e a abordagem contrafactual no Capítulo 5, este capítulo formaliza o conceito e o operacionaliza como um componente estrutural do pipeline do UrbanCounterViz.

Formalmente, seja $\mathcal{I} = \{I_1, I_2, \ldots, I_n\}$ um conjunto de SVI, cada uma associada a um escore de percepção $q_i \in [0, 10]$ estimado a partir de comparações pareadas (por exemplo, via Elo, TrueSkill ou Strength of Schedule, como descrito na Seção 2.4.2). Seja $\phi : \mathcal{I} \to \mathbb{R}^d$ uma função de embedding de características que mapeia cada imagem para uma representação d-dimensional. Um **par discordante** (I_i, I_j) é definido como um par de imagens que satisfaz duas condições simultaneamente:

- **Similaridade visual**: as imagens estão próximas no espaço de embedding, isto é, $sim(\phi(I_i), \phi(I_j)) \geq \tau$ para algum $\tau > 0$;
- **Divergência perceptual**: seus escores de percepção diferem substancialmente, isto é, $|q_i - q_j| \geq \sigma$ para algum $\sigma > 0$.

Os valores concretos usados neste trabalho são reportados na Seção 6.7. Intuitivamente, os pares discordantes são cenas que se parecem, mas que são julgadas de forma muito diferente por observadores humanos. Como as duas imagens compartilham a maior parte de seu conteúdo visual, qualquer diferença na percepção pode ser atribuída a um conjunto relativamente pequeno de elementos distintivos, em vez de à variação global da cena. Essa propriedade torna os pares discordantes particularmente informativos para a análise contrafactual: ao examinar sistematicamente as diferenças semânticas e em nível de objeto dentro de tais pares, é possível identificar quais elementos visuais — como a presença de vegetação, as condições de iluminação, as fachadas dos edifícios ou os sinais de desordem física — estão mais fortemente associados a deslocamentos na segurança percebida.

Enquanto as explicações contrafactuais perguntam *qual a mudança mínima em uma entrada que alteraria a predição de um modelo*, os pares discordantes fundamentam essa questão em dados reais observados: em vez de sintetizar modificações hipotéticas, a análise é conduzida por diferenças que de fato existem entre cenas urbanas reais. As intervenções resultantes são, portanto, mais plausíveis e diretamente interpretáveis como condições urbanas realistas. Essa estratégia contrafactual fundamentada constitui uma das contribuições metodológicas centrais do presente trabalho e é discutida em maior detalhe na Seção 6.5.

6.4.4 Construção do Grafo de Similaridade

Para capturar as relações visuais entre imagens em escala, um grafo de similaridade não direcionado $G = (\mathcal{V}, \mathcal{E})$ é construído, onde cada nó $v \in \mathcal{V}$ corresponde a uma SVI e cada aresta $(v_i, v_j) \in \mathcal{E}$ conecta imagens que são visualmente similares. Para cada imagem, um vetor de características profundas é extraído usando um modelo ResNet101 pré-treinado no ImageNet, sem ajuste fino sobre os escores de qualidade perceptual, garantindo que as representações de características capturem padrões visuais gerais independentemente dos escores. Isso mantém a topologia do grafo e as diferenças de escore como duas fontes de informação desacopladas, o que é essencial para identificar significativamente os pares discordantes. As similaridades de cosseno pareadas são então computadas em toda a coleção, e duas imagens são conectadas por uma aresta sempre que sua similaridade de cosseno excede um limiar τ, retendo apenas os pares fortemente similares.

Esse grafo formaliza o conceito de pares discordantes introduzido na Seção 6.4.3: pares de nós que estão próximos no espaço de embedding (conectados por uma aresta ou alcançáveis por um caminho curto), mas que diferem substancialmente em seus escores de percepção de segurança. Dentro do UrbanCounterViz, o grafo serve a dois propósitos complementares. Primeiro, ele fornece a base estrutural para a visão do *Seletor de Destino*, que apresenta imagens-alvo candidatas alcançáveis a partir de uma imagem de origem selecionada. Segundo, e mais centralmente, ele define os caminhos ao longo dos quais as edições contrafactuais são computadas: o caminho mais curto entre um nó de origem e um nó de destino com um escore de percepção de segurança mais alto decompõe uma lacuna perceptual potencialmente grande em uma sequência de transições menores e visualmente fundamentadas, cada uma informada por diferenças observadas entre cenas adjacentes reais.

Ao integrar escores de percepção, resumos semânticos e similaridade visual baseada em grafo, o estágio de pré-processamento produz as representações que possibilitam toda a exploração interativa e a geração contrafactual fundamentada subsequentes no UrbanCounterViz.

6.5 Pipeline de Edição Contrafactual Guiada por Caminho

O pipeline de edição guiada por caminho é o núcleo computacional do UrbanCounterViz. Dada uma imagem de origem I_{src} selecionada e uma imagem de destino I_{des}, o objetivo é gerar uma imagem contrafactual fundamentada que faça a origem transitar em direção à composição semântica do destino, preservando a coerência estrutural e a aparência fotorrealista da cena original. Isso é alcançado por meio de três componentes sequenciais: guia por caminho, modelagem de diferenças semânticas e edição de imagem orientada por VLM em dois estágios.

6.5.1 Guia por Caminho

Para um par selecionado (I_{src}, I_{des}), o caminho mais curto é computado entre seus nós correspondentes no grafo de similaridade definido na Seção 6.4.4. Esse caminho produz uma sequência de imagens intermediárias que são, cada uma, visualmente similares a seus vizinhos e que, coletivamente, fazem a ponte entre a origem e o destino:

$$I_{src} \to I_a \to I_b \to \cdots \to I_{des}.$$

O caminho é central para a interpretabilidade da abordagem. Em vez de exigir que o modelo de edição transite diretamente da origem para um alvo potencialmente distante — o que arriscaria produzir modificações irrealistas ou arbitrárias — a transformação é decomposta em etapas menores e contextualmente fundamentadas. Cada etapa é guiada

por diferenças de fato observadas entre cenas urbanas reais próximas, e não por uma otimização irrestrita do modelo ou por *prompting* de forma livre. Essa decomposição conecta-se diretamente à limitação-chave identificada no Capítulo 5: o UrbanCFX produzia contrafactuais no espaço de características que não podiam ser validados visualmente. Aqui, cada etapa ao longo do caminho é ela própria fundamentada em uma imagem real, garantindo que cada edição gerada esteja ancorada em uma condição urbana existente.

A geração procede sequencialmente ao longo do caminho: a cada etapa, a imagem atual é editada para produzir uma imagem atualizada que serve como origem para a etapa seguinte. As edições semânticas, portanto, acumulam-se incrementalmente, e o analista pode inspecionar a transformação em qualquer ponto intermediário, em vez de examinar apenas a saída final.

6.5.2 Modelagem de Diferenças Semânticas

A cada etapa do caminho, a composição semântica da imagem de origem atual é comparada com a da próxima imagem-alvo usando as proporções de categorias definidas na Seção 6.4.2. Seja $\hat{I}_{\mathcal{H}}$ a imagem de origem atual e X a próxima imagem-alvo ao longo do caminho. Para cada categoria semântica k, a diferença bruta com sinal é computada como:

$$\Delta_k = p_k(X) - p_k(\hat{I}_{\mathcal{H}}),$$

e normalizada como:

$$\delta_k = \frac{\Delta_k}{\max\Big(p_k(\hat{I}_{\mathcal{H}}),\, p_k(X)\Big)},$$

resultando em $\delta_k \in [-1, +1]$, onde valores positivos indicam categorias que devem aumentar, valores negativos indicam categorias que devem diminuir e valores próximos de zero indicam pouca ou nenhuma mudança. O vetor completo de diferença semântica é então:

$$\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_K).$$

Nem toda diferença semântica é igualmente significativa para a edição. Apenas as categorias com $|\delta_k|$ acima de um limiar mínimo são retidas e tratadas como categorias acionáveis para a etapa atual. Essa filtragem serve a dois propósitos: reduz o ruído das flutuações composicionais menores entre cenas similares e foca as ações de edição — e a atenção do analista — nas mudanças com maior probabilidade de produzir um deslocamento perceptual significativo. O vetor de diferença semântica filtrado $\boldsymbol{\delta}$ forma a entrada direta para o estágio de edição de imagem e também é exibido na interface para apoiar a transparência e a supervisão humana (R2, R3).

6.5.3 Edição de Imagem em Dois Estágios Orientada por VLM

A geração de imagens contrafactuais é implementada por meio de um fluxo de trabalho vision-language de dois estágios que separa *o que deve mudar* de *como a imagem*

é editada. Essa separação é uma decisão de projeto intencional que aprimora a interpretabilidade e a transparência: ao produzir uma representação intermediária explícita da transformação pretendida, o pipeline permite que os analistas inspecionem, compreendam e, se necessário, questionem a lógica da edição antes de examinar a imagem gerada.

Estágio 1 — Geração de Ações de Edição. No primeiro estágio, um conjunto de ações de edição candidatas é derivado do vetor de diferença semântica, da imagem de origem atual e dos escores de percepção de segurança dos estados atual e alvo. Um VLM processa a imagem de origem juntamente com uma orientação semântica estruturada e retorna um conjunto de ações de edição expressas como instruções visuais explícitas — por exemplo, *aumentar a cobertura de árvores ao longo da calçada*, *reduzir a proeminência dos cabos aéreos* ou *introduzir marcações de pavimento conservadas*. Essas ações são delimitadas pelo vetor de diferença semântica e formalmente definidas como:

$$\mathcal{E}_{\mathcal{H}}^{X} = \mathrm{VLM}_1\Big(\mathcal{P}_{actions}(\hat{I}_{\mathcal{H}}, \boldsymbol{\delta}, Q_{\hat{I}_{\mathcal{H}}}^{\mathrm{safety}}, Q_{X}^{\mathrm{safety}})\Big)\,.$$

O conjunto de ações resultante é exibido na *Visão de Geração* da interface, onde os analistas podem revisá-lo antes de inspecionar a imagem gerada. O Prompt 2 detalha todas as instruções, o contexto e o fluxo de raciocínio.

```
[System] You are generating targeted edit instructions for a pair of urban street view images,
     using semantic segmentation data that expresses each scene element as a proportion of
    total pixel area (%).
Categories: {list_categories}

---
## EDIT DIRECTION: {maintenance_state}
- **well-maintained** -> elements showing evidence of active and maintenance.
- **deteriorated** -> elements showing evidence of wear, damage, or disuse.
Apply this direction by choosing the *condition state* of elements - not which specific
    objects to place or remove. Select the element type that is most plausible given the
    existing scene; then describe its condition according to the direction. All modifications
    should be smooth not excessive.

---
## CATEGORIES TO ACTIONS
Pre-selected where |difference| > 0.10. Each entry shows the category's current share and the
    pixel impact to achieve.
{category_diff_list}
{{category}} (currently {{current_pct}}\%): sign {{impact_pct}}\%; If sign is "+" = this
    category gains that pixel share
  else this category loses that pixel share

---
## SPATIAL REASONING - DO THIS BEFORE WRITING ANY ACTION
Before writing each action, answer internally:
1. **What does this category currently occupy in the scene?**
   Infer from its current percentage and the other categories present.
   (e.g., if 'floor' is 10\% and 'terrain_vehicle' is 25\%, most floor is likely occluded by
    vehicles)
2. **Where in the frame would this action plausibly occur?**
   (foreground, mid-distance, far background; left / center / right)
   Ground-level elements occupy foreground-mid; sky and upper facade occupy top of frame.
3. **Does this action conflict with any other simultaneous action?**
   Two categories increasing cannot compete for the same spatial zone.
   If a conflict exists, resolve by spatial separation or assign the contested zone to the
    category with the larger impact value.
4. **Is this action physically plausible in a street scene?**
   The element type must be consistent with what could exist in this spatial locationin an
    ordinary urban environment. Do not introduce elements that would require implausible
    spatial conditions.
Only after this reasoning, write the action.
---
```

Prompt 2 – *Prompt* utilizado para gerar uma lista de ações

```
##  CATEGORY RULES
| Category | Scope | Maintenance direction applies? |
|---|---|---|
| vegetation | Plants, shrubs, trees, soil patches visible in frame | Yes |
| construction | Building facades, windows, doors, surface coverings | Yes |
| floor | Pavement, markings, ground-level objects | Yes |
| sky | Roof elements or overhangs that occlude sky | No |
| city_elements | Standard urban infrastructure: lighting, signage, street furniture. No
    surveillance, security, or enforcement elements. | Yes |
| terrain_vehicle | Standard civilian vehicles only. No damaged, burning, or abandoned
    vehicles. | No |
| human | Pedestrians or cyclists only. Max 2 individuals. Describe only activity and distance.
     No appearance, clothing, age, gender, or group composition. | No |

---
## ACTION RULES
1. **One action per category** - either add or remove. No repositioning, resizing, or
    replacement.
2. **One sentence.** Direct and concrete. No qualifiers, hedges, or percentages.
3. **For additions:** specify element type, condition state (if applicable), scene location,
    and approximate distance (close  middistance  far).
4. **For removals:** specify what is removed and from where. Do not name what is revealed
    unless it is the explicit purpose of another simultaneous action.
5. **No action may modify an unlisted category as a side effect.** If a side effect is
    unavoidable, do not perform that action.
6. **Do not name specific brand, model, or culturally specific objects.** Use generic element
    types.
7. **Do not use examplederived element names.** Choose element types from the scene context,
    not from memory of similar instructions.

---
## OUTPUT - strict JSON only, no markdown:
{{
  "action": {{
    "<category>": "<single action sentence>"
  }}
\text{}}}

Only include categories listed in CATEGORIES TO ACTIONS.
```

Prompt 2 – (continuação)

Estágio 2 — Edição Fotorrealista de Imagem. No segundo estágio, as ações de edição $\mathcal{E}_{\mathcal{H}}^{X}$ e a imagem de origem atual são passadas a um segundo VLM, que produz a imagem editada como:

$$\hat{I}_{\mathcal{H}}^{X} = \mathrm{VLM}_2\Big(\mathcal{P}_{image}(\hat{I}_{\mathcal{H}}, \mathcal{E}_{\mathcal{H}}^{X})\Big)\,,$$

aplicando as modificações solicitadas enquanto preserva o conteúdo restante da cena, a estrutura espacial, o ponto de vista e a coerência visual geral da cena original. O Prompt 3 detalha as restrições e regras para editar imagens, visando manter as características visuais originais.

```
[System] You are a photo editor working on an urban scene image.

Given the original image and the following editing actions by visual category:
{actions}

**ABSOLUTE CONSTRAINTS:**
- **No geometry changes.** Do not alter camera angle, perspective, vanishing points, or the
    perceived size of any element. All edits must be flat and in-plane.
- **No category contamination.** Do not modify any element from a category not listed above -
    not even incidentally. If an unlisted category is accidentally affected, revert it.
- **Preserve surface detail.** Retain all textures, graffiti, weathering, and signage on
    existing surfaces exactly as they appear, unless explicitly targeted by an action.
- **Preserve image style.** Keep the original color grading, contrast, and photographic style
    intact throughout. Do not enhance or stylize any part of the image.

**EDITING RULES:**
- Execute each action literally - removals only remove, additions only add.
  Never compensate a removal by adding elements from another category.
- **On removal:** fill the exposed area by extending the immediately surrounding surface (
    pavement, wall, sky) using adjacent pixels as reference. If the removed element was in the
     foreground, reconstruct the background layer behind it at the correct depth. Continue any
     texture or graffiti present on the fill surface - do not leave a clean patch. Never
    introduce a new category as fill.
- **On addition:** place the element at a spatially natural location for its category. Match
    its scale to the depth plane, replicate the scene's light direction and shadows, and match
     the photographic style (color temperature, contrast, sharpness). Do not occlude elements
    from unlisted categories.
- **If an action is impossible** (element not present, no valid surface), skip it silently.

Return only the edited image.
```

Prompt 3 – *Prompt* utilizado para editar uma imagem de entrada seguindo uma lista de ações

Após a geração, a imagem editada é re-segmentada com o OneFormer, e suas proporções de categorias semânticas, valores de indicadores urbanos e escore de segurança estimado são recomputados. Essas saídas derivadas permitem que a interface compare imediatamente a imagem gerada com a origem e o destino ao longo de múltiplas dimensões (R4), possibilitando que os analistas avaliem se a edição moveu a cena na direção perceptual pretendida e se as mudanças são visualmente coerentes e semanticamente consistentes com as diferenças orientadoras.

Esse projeto de dois estágios aprimora a interpretabilidade ao tornar o processo de edição transparente e auditável. Diferentemente das abordagens que geram diretamente uma imagem modificada, o UrbanCounterViz separa explicitamente a identificação do que deve mudar do ato de mudá-lo, garantindo que cada imagem gerada seja acompanhada das diferenças semânticas que a motivaram, das ações de edição que a guiaram e dos resumos perceptuais e semânticos recomputados que a avaliam.

6.6 A Interface do UrbanCounterViz

O UrbanCounterViz é um sistema interativo de análise visual que integra o pipeline computacional descrito acima em um ambiente exploratório unificado. A interface compreende seis visões coordenadas — *Seletor de Origem*, *Seletor de Destino*, *Visão de Caminho*, *Visão de Imagem*, *Visão de Geração* e *Estatísticas da Imagem* — que, coletivamente, apoiam as tarefas analíticas definidas na Seção 6.3.2. A Figura 21 apresenta a interface completa; a Tabela 11 resume o mapeamento entre as visões e as tarefas.

Table 11 – Visões do UrbanCounterViz e as tarefas analíticas que elas apoiam.

View	***T1***	***T2***	***T3***
Source Selector	✓		
Destination Selector	✓		
Path View	✓	✓	✓
Image View		✓	✓
Generation View		✓	✓
Image Stats			✓

6.6.1 Seletor de Origem (A)

O *Seletor de Origem* é o ponto de entrada para o fluxo de trabalho analítico. Ele apresenta uma visão geral baseada em mapa da coleção de imagens (Figura 21A), onde cada ponto representa uma SVI geolocalizada colorida por seu Q-score perceptual. Essa codificação espacial permite que os analistas identifiquem imediatamente áreas de baixa, moderada ou alta segurança percebida e contextualizem sua análise dentro da geografia da cidade.

Selecionar um ponto designa a imagem de origem I_{src} a ser transformada e inicializa o fluxo de trabalho subsequente: o *Seletor de Destino* é atualizado para mostrar os alvos candidatos alcançáveis a partir de I_{src} dentro do grafo de similaridade, e todas as outras visões são reiniciadas para refletir a nova seleção. Essa visão apoia primariamente a T1 ao fornecer a fundamentação espacial e perceptual que orienta o analista antes que qualquer raciocínio contrafactual comece.

Figure 21 – Interface do UrbanCounterViz. (A) **Seletor de Origem**: imagens geolocalizadas coloridas pelo Q-score perceptual; a seleção de uma origem inicializa o fluxo de trabalho. (B) **Seletor de Destino**: gráfico de dispersão das imagens candidatas mostrando a similaridade visual em relação à origem (x) versus o Q-score (y), apoiando a seleção do destino. (C) **Visão de Caminho**: caminho mais curto no grafo de similaridade da origem ao destino, expondo cenas intermediárias e suas diferenças semânticas. (D) **Visão de Imagem**: comparação lado a lado da origem (D.1) e do contrafactual gerado (D.2), acrescida de sobreposições que destacam as regiões adicionadas e removidas. (E) **Visão de Geração**: ações de edição passo a passo organizadas por categoria semântica, permitindo a inspeção do processo de geração. (F) **Estatísticas da Imagem**: resumos coordenados que comparam a composição semântica (F.1) e os indicadores urbanos (F.2) entre as imagens de origem, de destino e geradas. Fonte: o autor.

6.6.2 Seletor de Destino (B)

O *Seletor de Destino* apresenta um gráfico de dispersão das imagens candidatas alcançáveis a partir de I_{src} dentro do grafo de similaridade (Figura 21B). O eixo horizontal codifica a similaridade visual em relação à imagem de origem e o eixo vertical codifica o Q-score perceptual, criando um espaço bidimensional no qual o analista pode simultaneamente raciocinar sobre a relação entre cenas e a distância perceptual. A região superior direita desse gráfico de dispersão é de particular interesse analítico: as imagens ali localizadas são tanto visualmente similares à origem quanto associadas a uma segurança percebida substancialmente mais alta — precisamente as condições que definem um par discordante forte e um candidato de alta qualidade para a edição contrafactual.

Ao tornar o compromisso entre similaridade e percepção de segurança visualmente explícito, essa visão apoia a T1 e dá ao analista controle direto sobre o grau de desvio visual que o contrafactual resultante envolverá. Selecionar um ponto no gráfico de dispersão designa I_{des} e dispara o cálculo do caminho e as atualizações da interface em todas as visões subsequentes.

6.6.3 Visão de Caminho (C)

A *Visão de Caminho* visualiza o caminho mais curto no grafo de similaridade entre I_{src} e I_{des} (Figura 21C), conforme computado na Seção 6.5.1. Ela expõe as imagens intermediárias que conectam a origem e o destino por meio de cenas visualmente relacionadas, renderizando o processo de transformação como uma sequência de transições incrementais, em vez de uma única mudança abrupta. Os nós são posicionados ao longo do eixo horizontal pela dissimilaridade pareada e ao longo do eixo vertical pelo Q-score perceptual, possibilitando a inspeção de se o caminho segue uma trajetória perceptual suave ou envolve mudanças de escore não monotônicas. As diferenças semânticas entre nós consecutivos são retratadas como gráficos de barras compactos que resumem as mudanças de categoria mais significativas em cada etapa.

Essa visão é analiticamente central: ela apoia a T1 ao expor as relações estruturais entre cenas similares, a T2 ao evidenciar as diferenças semânticas passo a passo que guiarão a edição, e a T3 ao possibilitar que os analistas antecipem como os escores perceptuais mudam ao longo da trajetória antes que a geração comece. As seleções nessa visão propagam-se para a *Visão de Imagem* e as *Estatísticas da Imagem*, permitindo a inspeção focada de qualquer etapa individual.

6.6.4 Visão de Imagem (D)

A *Visão de Imagem* apoia a comparação visual direta entre a imagem de origem e o contrafactual gerado (Figura 21D). Por padrão, o painel esquerdo exibe I_{src} e o painel direito exibe a imagem gerada atual $\hat{I}$; as etapas intermediárias podem ser carregadas por meio de seleções na *Visão de Caminho* ou na *Visão de Geração*. Para facilitar a identificação de diferenças entre imagens visualmente similares — uma tarefa cognitivamente exigente quando as mudanças são sutis ou localizadas — o sistema sobrepõe máscaras de mudança semântica que distinguem as regiões adicionadas (categorias que aumentaram) das regiões removidas (categorias que diminuíram). O resultado funciona como um diff visual, permitindo que os analistas localizem eficientemente as modificações e avaliem se elas correspondem a mudanças estruturalmente plausíveis na cena.

Esse mecanismo de mascaramento é uma resposta direta ao Requisito R3: ele torna as edições explícitas e espacialmente localizadas, em vez de deixar que o analista as detecte por meio de uma inspeção visual não guiada. Juntamente com a *Visão de Geração*, ela apoia a T2 ao confirmar os efeitos espaciais do direcionamento semântico e a T3 ao possibilitar a interpretação detalhada de como as mudanças geradas são realizadas na imagem.

6.6.5 Visão de Geração (E)

A *Visão de Geração* apresenta o processo de edição contrafactual como uma sequência explícita de etapas de geração (Figura 21E). Para cada etapa, a visão exibe as ações de edição derivadas no Estágio 1 do pipeline VLM, organizadas por categoria semântica. Essa visão é central para a interpretabilidade do sistema: ao revelar as representações de ação intermediárias que guiam o modelo de edição de imagem, o UrbanCounterViz torna a geração transparente e auditável, em vez de tratá-la como uma caixa-preta. Os analistas podem revisar essas ações ao lado das imagens geradas correspondentes e avaliar se as intenções declaradas correspondem aos resultados visuais observados — ou se deveriam selecionar um caminho diferente, ajustar o destino ou considerar uma trajetória alternativa.

Essa representação explícita de ações também serve a um propósito prático identificado por especialistas de domínio em uso preliminar: ela permite que os analistas prevejam o caráter de uma edição antes de examinar a imagem gerada, reduzindo a surpresa cognitiva e tornando o comportamento do sistema mais confiável. A *Visão de Geração* apoia primariamente a T2 (compreender como as diferenças semânticas se traduzem em instruções de edição) e a T3 (auditar a relação entre as mudanças pretendidas e os resultados observados). As seleções nessa visão atualizam tanto a *Visão de Imagem* quanto as *Estatísticas da Imagem* para a inspeção em nível de etapa.

6.6.6 Estatísticas da Imagem (F)

A visão de *Estatísticas da Imagem* fornece um resumo multidimensional do resultado da edição usando dois painéis coordenados (Figura 21F) que comparam simultaneamente as imagens de origem, de destino e geradas.

Painel de Composição Semântica (F.1). Este painel compara as proporções agregadas de categorias semânticas das três imagens. Ele possibilita que os analistas avaliem se a imagem gerada moveu-se em direção ao destino em termos de composição da cena — por exemplo, se a vegetação aumentou ou se os elementos de construção mudaram — e se os deslocamentos observados correspondem às diferenças semânticas que motivaram a edição.

Painel de Indicadores Urbanos (F.2). Este painel compara os seis indicadores de percepção urbana de nível mais alto, juntamente com o escore de segurança estimado, entre as imagens de origem, de destino e gerada. Os eixos são normalizados por dimensão para facilitar a comparação entre medidas heterogêneas. Este painel permite que os analistas interpretem se as mudanças visuais na imagem gerada correspondem a deslocamentos significativos em qualidades ambientais como openness, greenness, walkability e enclosure — dimensões que importam para as decisões de projeto urbano para além do escore de percepção isoladamente.

Juntos, esses painéis implementam o Requisito R4 ao conectar as edições visíveis

da cena às mudanças na composição semântica, nos indicadores urbanos e na segurança percebida estimada. Essa visão multidimensional também serve como uma verificação de qualidade: se a imagem gerada produzir deslocamentos inesperados em indicadores não visados pela edição (por exemplo, uma grande diminuição em openness quando apenas a vegetação deveria aumentar), o analista pode sinalizar isso como um indício de instabilidade da geração e selecionar uma etapa alternativa.

6.7 Detalhes de Implementação

O UrbanCounterViz emprega uma arquitetura cliente–servidor. O cliente é implementado em JavaScript usando o D3.js (Bostock et al., 2011) para gráficos interativos e interações coordenadas em todas as seis visões. O servidor é implementado em Python usando o FastAPI (Ramírez, 2019), que serve os dados pré-processados, as informações do grafo, os ativos de imagem e as saídas de geração.

As características visuais profundas para a construção do grafo de similaridade são extraídas usando um modelo ResNet101 pré-treinado no ImageNet (He et al., 2016). A segmentação semântica é realizada com o OneFormer pré-treinado no ADE20K (Jain et al., 2023), acessado por meio do ecossistema HuggingFace (Wolf et al., 2019). A estimativa do escore de percepção de segurança após a geração é conduzida usando um modelo ConvNext treinado no conjunto de dados Place Pulse. A edição de imagem usa o Gemini `2.5-flash` e o `2.5-flash-image` (Google, 2025) para o Estágio 1 (geração de ações) e o Estágio 2 (edição fotorrealista), respectivamente. Todas as etapas de pré-processamento — incluindo extração de características, segmentação, cálculo de escores e construção do grafo — são realizadas offline para garantir interações responsivas durante a análise ao vivo. Apenas a geração de imagens é disparada sob demanda durante uma sessão.

Os limiares que governam o pipeline — similaridade visual (τ), divergência perceptual (σ) e filtragem de diferença semântica — controlam o compromisso entre fundamentação e diversidade. Valores mais estritos produzem pares mais visualmente similares e mais fáceis de interpretar, mas reduzem o número de transformações disponíveis; valores mais frouxos expandem o espaço de busca ao custo de edições menos plausíveis. Neste trabalho, esses parâmetros foram selecionados empiricamente para favorecer a interpretabilidade e a coerência visual em detrimento da cobertura exaustiva: por exemplo, $\tau = 0.85$ para a criação de arestas, $\sigma = 4.0$ pontos de Q-score para a divergência perceptual e $|\delta_k| > 0.10$ para a filtragem de diferença semântica.

6.8 Considerações Finais

Este capítulo introduziu o UrbanCounterViz, um sistema de análise visual que aborda as principais limitações identificadas ao longo dos capítulos anteriores desta

tese. Onde o Capítulo 4 produziu atribuições espaciais estáticas e o Capítulo 5 produziu recomendações apenas textuais no espaço de características, o UrbanCounterViz gera transformações de cenas urbanas fotorrealistas, fundamentadas e interativamente exploráveis. Ao fundamentar cada edição em diferenças observadas entre cenas adjacentes reais em um grafo de similaridade, o sistema evita a principal fraqueza das abordagens generativas anteriores — a produção de modificações visualmente impressionantes, mas analiticamente ininterpretáveis. Ao tornar todo o processo de transformação — da seleção de caminho ao cálculo de diferenças semânticas, à geração de ações de edição, à saída fotorrealista — transparente e inspecionável dentro de uma interface coordenada de análise visual, ele apoia o tipo de raciocínio com humano no processo que as decisões de planejamento urbano exigem.

Duas propriedades do projeto merecem ênfase, pois distinguem o UrbanCounterViz de todos os trabalhos relacionados anteriores revisados neste capítulo. Primeiro, a fundamentação: as intervenções são derivadas de padrões que existem em dados reais, e não de otimização do modelo ou de *prompting* irrestrito, tornando-as mais críveis como evidência para decisões de projeto. Segundo, a transparência: cada estágio do pipeline é exposto ao analista por meio da interface, desde o grafo de similaridade e os destinos candidatos até os vetores de diferença semântica, as ações de edição, as imagens geradas e os resumos multidimensionais de resultados. Juntas, essas propriedades apoiam não apenas a geração de cenas urbanas contrafactuais, mas o raciocínio interpretável e fundamentado em evidências sobre essas cenas, que constitui a contribuição central desta tese.

O Capítulo 7 apresenta a avaliação do UrbanCounterViz por meio de três cenários de uso, um estudo de feedback com especialistas e um estudo de feedback com usuários, examinando como o sistema apoia a análise em diferentes condições urbanas, comprimentos de trajetória e direções de transformação.

7 Avaliação do UrbanCounterViz

O capítulo anterior descreveu como o UrbanCounterViz transforma o raciocínio construído ao longo dos últimos três capítulos em algo com que um analista pode de fato trabalhar: uma cena, uma edição plausível e uma forma de vê-la. O que permanecia por testar era se ele de fato funciona como pretendido para as pessoas que deveriam usá-lo. Este capítulo faz essa pergunta diretamente, pelos olhos das pessoas que usariam tal sistema — planejadores, analistas e observadores cotidianos de uma rua.

A avaliação está organizada em três partes complementares, cada uma abordando uma dimensão diferente da questão avaliativa central: o UrbanCounterViz apoia o raciocínio transparente, com humano no processo, sobre transformações urbanas plausíveis de maneiras que nem os sistemas de análise visual anteriores nem os métodos generativos anteriores forneceram? A primeira parte apresenta três cenários de uso conduzidos com SVI de São Paulo, cada um visando um objetivo analítico diferente: transformar uma cena de baixa segurança em uma visualmente similar, porém de maior segurança; explorar o comportamento de trajetórias mais longas de múltiplas etapas; e examinar a geração contrafactual na direção reversa, rumo a uma menor segurança percebida. Esses cenários não pretendem ser experimentos formais com usuários, mas demonstrações estruturadas do alcance analítico do sistema — o que ele revela, onde se destaca e onde suas saídas se tornam menos confiáveis. A segunda parte reporta o feedback qualitativo coletado por meio de entrevistas semiestruturadas com especialistas de domínio em planejamento urbano e análise urbana, fornecendo uma perspectiva externa sobre a usabilidade, a utilidade e a interpretabilidade do sistema a partir de profissionais capazes de julgar o realismo e a relevância prática das edições geradas. A terceira parte reporta um estudo estruturado com usuários no qual uma amostra mais ampla de participantes avaliou o sistema usando instrumentos de usabilidade padronizados e questões de percepção direcionadas, fornecendo evidência quantitativa que complementa o julgamento qualitativo dos especialistas.

Juntas, as três partes abordam essa questão a partir de ângulos cada vez mais amplos: os cenários demonstram o que o sistema pode fazer, as entrevistas com especialistas testam se os profissionais de domínio o consideram crível, e o estudo com usuários verifica se essa credibilidade se sustenta para um público mais amplo e não especialista.

7.1 Cenários de Uso

Os cenários de uso empregam SVI de São Paulo, Brasil, selecionadas do conjunto de dados Place Pulse 2.0. São Paulo fornece uma coleção grande e visualmente diversa de cenas urbanas que abrange uma ampla gama de escores de segurança percebida, condições

de infraestrutura e composições semânticas — tornando-a bem adequada para demonstrar o alcance analítico do sistema em diferentes tipos de ambientes urbanos. Três cenários são apresentados em ordem crescente de complexidade: uma transformação focada de duas etapas, uma trajetória monotônica mais longa e uma trajetória na direção perceptual reversa.

7.1.1 Cenário 1 — Aprimorando a Percepção de Segurança Urbana por meio de Pares Discordantes

Figure 22 – Cenário de uso mostrando uma transformação contrafactual guiada por caminho, de uma cena urbana de menor segurança para uma mais segura. Partindo de uma imagem de origem de baixa segurança S (escore: 1,58) e de uma imagem de destino visualmente semelhante, porém de maior segurança, D (escore: 6,11, similaridade: 0,80), o UrbanCounterViz calcula um caminho com um nó intermediário (I.1) e gera duas edições fundamentadas (G.1 e G.2). Os *prompts* de ação correspondentes (P.1 e P.2) e as máscaras semânticas (M.C e M.V) ajudam a explicar como a transformação se desenrola. Fonte: o autor.

Este cenário ilustra o objetivo analítico primário do UrbanCounterViz: transformar uma cena urbana de baixa segurança em uma alternativa visualmente similar, porém de maior segurança, guiada por diferenças observadas em um par discordante real.

Configuração. O analista começa no *Seletor de Origem* identificando uma imagem de baixa pontuação em uma área residencial de São Paulo (Figura 22S, Q-score: 1,58). A cena mostra uma rua estreita com vegetação limitada, mobiliário urbano esparso e poucos sinais visíveis de atividade — características visuais consistentemente associadas a uma menor segurança percebida nas análises dos Capítulos 4 e 5. No *Seletor de Destino*, o analista identifica uma imagem visualmente similar com segurança percebida substancialmente mais alta (Figura 22D, Q-score: 6,11, similaridade: 0,80). Esse par satisfaz a definição formal de par discordante introduzida na Seção 6.4.3: duas cenas que estão próximas no

espaço de embedding visual, mas que diferem fortemente em seus escores de segurança atribuídos por humanos.

Geração de caminho e de edições. A *Visão de Caminho* revela um caminho curto com um nó intermediário (Figura 22I.1), decompondo a transformação em duas etapas de edição. Para cada etapa, o pipeline deriva ações de edição explícitas (Figura 22P.1 e P.2) a partir das diferenças semânticas computadas entre nós adjacentes ao longo do caminho. Essas ações incluem aumentar a cobertura de árvores ao longo da calçada, modificar estruturas visíveis à margem da rua e introduzir pistas associadas a ambientes mais bem conservados e mais ativos. As diferenças de categoria semântica que impulsionam cada etapa são exibidas na *Visão de Caminho* como gráficos de barras compactos (Figura 22T.1), dando ao analista uma prévia clara do que mudará antes de examinar as imagens geradas.

Resultados e interpretação. A imagem gerada final (Figura 22G.2) aumenta o escore estimado de percepção de segurança em aproximadamente 4,5 pontos em relação à origem, preservando o arranjo original da rua, a estrutura dos edifícios e o ponto de vista. As máscaras de mudança semântica (Figura 22M.C e M.V) destacam as modificações mais relevantes nas categorias *construction* e *vegetation*, possibilitando que o analista verifique que as mudanças são espacialmente localizadas e estruturalmente coerentes, em vez de globalmente inconsistentes. A visão de *Estatísticas da Imagem* confirma que a imagem gerada move-se em direção ao destino tanto na composição semântica quanto nos indicadores urbanos, incluindo greenness e openness.

Este cenário demonstra uma propriedade-chave da abordagem contrafactual fundamentada: em vez de estimular um modelo generativo com uma descrição de forma livre de uma cena desejada, as edições são derivadas de diferenças de fato observadas entre dois ambientes urbanos reais. O resultado não é apenas visualmente mais coerente, mas também analiticamente mais crível — o analista pode rastrear exatamente quais diferenças no par real motivaram cada edição e pode inspecionar se a imagem gerada realiza essas diferenças fielmente.

7.1.2 Cenário 2 — Análise de Trajetória de Múltiplas Etapas

Este cenário explora o comportamento do UrbanCounterViz ao longo de trajetórias contrafactuais mais longas, examinando tanto o que se ganha quanto o que se perde à medida que o número de etapas de edição aumenta.

Configuração. Partindo da mesma imagem de origem de baixa pontuação do Cenário 1, o analista seleciona um destino perceptualmente mais distante (Figura 21C, Q-score: 8,0, similaridade: 0,71) — colocando-o entre as cenas de mais alta pontuação no subconjunto de São Paulo. Um caminho monotônico é então identificado, exigindo que o escore de segurança percebida aumente a cada etapa. A trajetória resultante passa por seis nós e

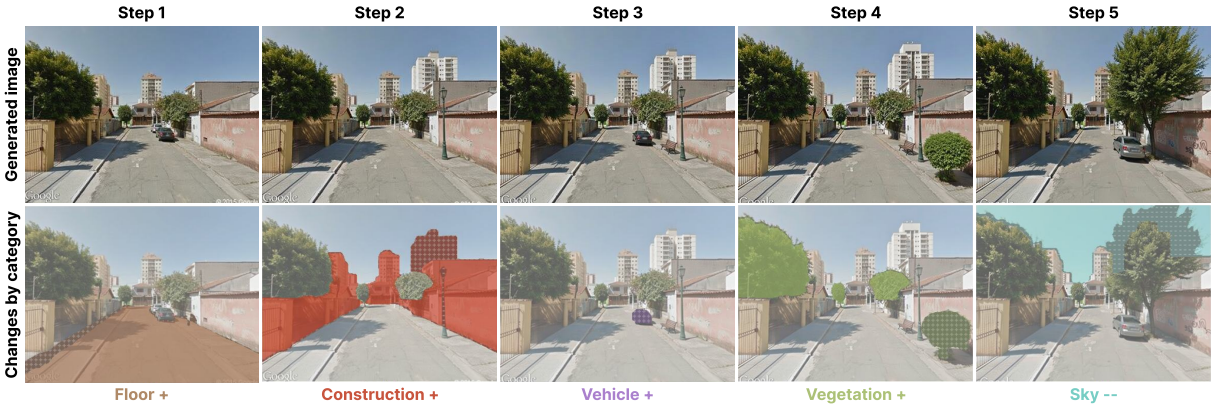

Figure 23 – Cenário de uso mostrando uma trajetória contrafactual de múltiplas etapas em direção a uma maior segurança percebida. Cada coluna corresponde a uma etapa de geração, com a imagem gerada na parte superior e a máscara de mudança semântica dominante abaixo. As regiões destacadas indicam a classe mais fortemente modificada em cada etapa; "+" denota um aumento na presença e "–" uma forte diminuição. A primeira etapa é calculada em relação à imagem de origem, e as etapas subsequentes são calculadas em relação à imagem gerada anteriormente. Fonte: o autor.

envolve cinco imagens geradas sequenciais (Figura 23).

Transformação progressiva. Ao longo da trajetória, o pipeline introduz progressivamente elementos comumente associados a uma maior segurança percebida: estrutura construída adicional, mobiliário urbano, vegetação expandida e mudanças no céu visível. A visão de *Estatísticas da Imagem* (Figura 21F) confirma que a imagem gerada final move-se substancialmente para mais perto do destino tanto na composição semântica quanto no escore de segurança estimado. Crucialmente, a trajetória de múltiplas etapas revela *como* essas mudanças se acumulam ao longo do tempo: o analista pode inspecionar cada transformação intermediária, em vez de apenas o resultado final, e pode identificar em qual etapa uma determinada categoria muda significativamente pela primeira vez.

Retornos decrescentes e modos de falha. A trajetória também revela limitações importantes da geração estendida. Como mostrado na Figura 23, os maiores ganhos ocorrem durante as duas ou três primeiras etapas. As edições iniciais introduzem modificações visualmente salientes e globalmente coerentes, como vegetação expandida, elementos construídos adicionados e um ambiente de rua mais estruturado. Nas etapas posteriores, o escore de segurança estimado continua a aumentar, mas os ganhos tornam-se menores e mais localizados. Dois modos de falha recorrentes emergem. Primeiro, a trajetória pode entrar em um **regime oscilatório**, no qual as edições em etapas posteriores sobrescrevem ou contrariam parcialmente as modificações introduzidas em etapas anteriores. Esse comportamento é visível em mudanças repetidas na categoria *vehicles* nas Etapas 3 e 5 (Figura 23), onde o pipeline alternadamente adiciona e remove elementos de veículos sem convergir para um estado semântico estável. Segundo, algumas edições tornam-

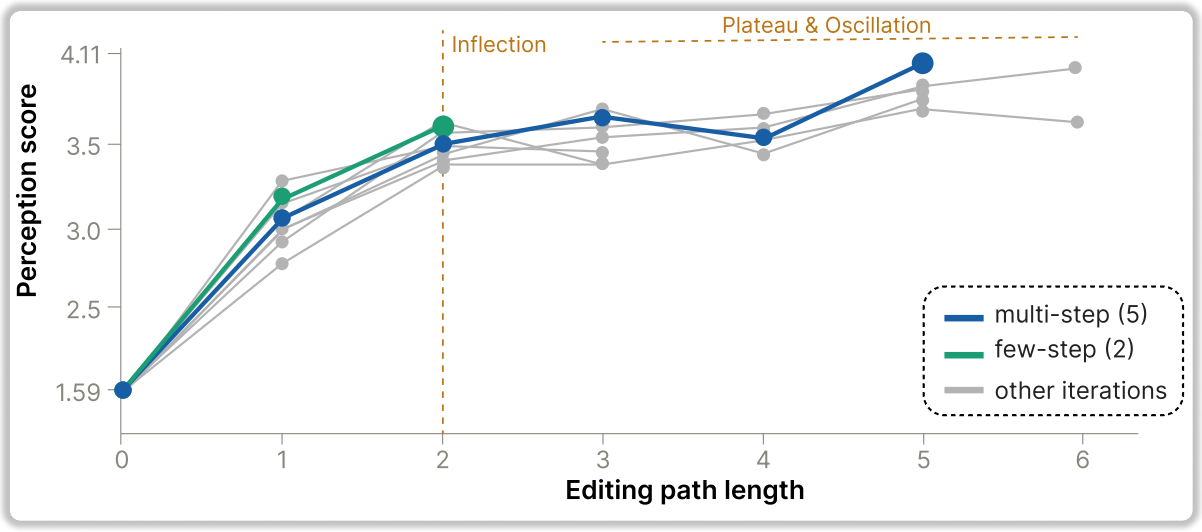

Figure 24 – Escore de percepção estimado em função do comprimento do caminho de edição para as trajetórias de múltiplas etapas (azul) e de poucas etapas (verde), juntamente com execuções adicionais a partir da mesma imagem de origem (cinza). Todas as execuções exibem um comportamento consistente: ganhos iniciais rápidos, uma inflexão em torno da etapa 2–3 e um platô com oscilação nas etapas posteriores. Fonte: o autor.

se geometricamente implausíveis como modificações localizadas: a inserção de grandes elementos estruturais que não se integram naturalmente ao arranjo da rua ao redor. Esses efeitos são tornados legíveis por meio das máscaras de mudança semântica e dos resumos de *Estatísticas da Imagem*, permitindo que o analista identifique o ponto de inflexão no qual a geração se torna não confiável.

Comparação com o Cenário 1. O contraste entre os dois cenários esclarece um compromisso fundamental. A edição de poucas etapas, como no Cenário 1, produz intervenções mais interpretáveis e visualmente plausíveis — adicionar vegetação, melhorar pistas de conservação, ajustar elementos construídos — que se alinham a ações realistas de projeto urbano. Trajetórias monotônicas mais longas expandem o espaço de escores perceptuais alcançáveis, mas introduzem uma inconsistência visual e uma instabilidade semântica crescentes. Para testar se essa observação é específica de um único caminho, trajetórias adicionais a partir da mesma imagem de origem foram computadas usando caminhos monotônicos alternativos rumo a destinos de similaridade comparativamente baixa (ver Figura 24). Ao longo dessas execuções, o padrão é consistente: ganhos rápidos nas etapas iniciais, um ponto de inflexão em torno das Etapas 2–3 e retornos decrescentes com comportamento oscilatório em seguida. Isso sugere que o compromisso entre trajetórias curtas e longas reflete uma propriedade estrutural do pipeline, e não um artefato de qualquer escolha de caminho isolada.

Esse achado destaca o valor prático do projeto com humano no processo do UrbanCounterViz: em vez de gerar a trajetória completa automaticamente e apresentar apenas

o ponto final, o sistema permite que os analistas inspecionem cada etapa, identifiquem o ponto de inflexão e interrompam ou redirecionem a trajetória quando as saídas começam a divergir de seus objetivos analíticos.

7.1.3 Cenário 3 — Diminuição Progressiva da Segurança Percebida

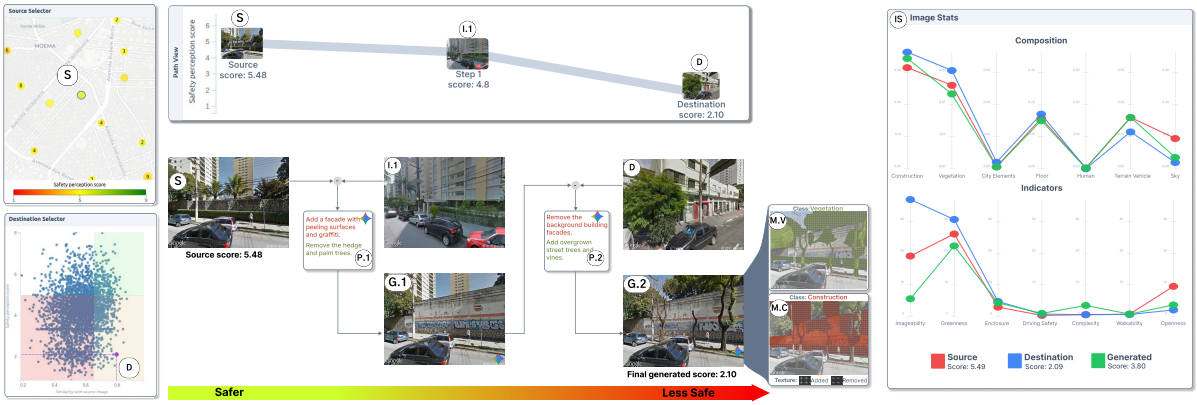

Figure 25 – Cenário de uso mostrando uma diminuição progressiva da segurança percebida, transformando uma imagem de aparência relativamente segura (S) em outra de aparência menos segura (G.2) por meio de um processo de geração de duas etapas. O método produz um resultado intermediário (G.1) que já aparenta estar mais negligenciado. A mudança mais proeminente é a transformação do muro decorado em um muro mais deteriorado coberto por pichações. O painel de Estatísticas da Imagem (IS) reflete essa redução em Vegetação e o aumento em Construção, que por sua vez diminuem os indicadores de *Imageability* e *Greenness*, ambos frequentemente associados à segurança percebida. Fonte: o autor.

O terceiro cenário examina a direção reversa do processo de transformação. Em vez de melhorar uma cena de baixa segurança, o analista usa o UrbanCounterViz para transformar progressivamente uma rua de aparência relativamente segura em outra percebida como menos segura. Essa capacidade bidirecional tem um propósito analítico distinto: permite que pesquisadores e planejadores identifiquem quais pistas visuais — quando introduzidas ou removidas — mais fortemente sinalizam desordem, negligência ou segurança reduzida a observadores humanos.

Configuração. A imagem de origem é uma cena com um escore de segurança médio-alto (Figura 25S, Q-score: 5,48), mostrando uma rua residencial conservada com vegetação visível e um muro decorado. O analista seleciona como destino uma imagem visualmente similar com um escore de segurança percebida substancialmente mais baixo (Figura 25D, Q-score: 2,10). O grafo de similaridade revela um caminho curto de duas etapas entre a origem e o destino, possibilitando uma trajetória reversa focada e interpretável.

Primeira etapa — introduzindo desordem estrutural. Na primeira etapa de edição, o pipeline gera um resultado intermediário (Figura 25G.1) transformando a origem em

direção à imagem de referência intermediária (Figura 25I.1). Como essa cena de referência é dominada por elementos relacionados à construção, as ações de edição correspondentes (Figura 25P.1) enfatizam o aumento da proeminência visual das estruturas construídas ao mesmo tempo em que reduzem a vegetação. As mudanças mais salientes na imagem gerada são a introdução de um muro deteriorado com tinta descascando e pichação, e a remoção da cerca viva e das palmeiras que caracterizavam a cena de origem. Essas modificações instanciam diretamente as pistas visuais identificadas no Capítulo 4 como associadas à percepção insegura e operacionalizam o vocabulário de elementos de desordem desenvolvido no Capítulo 5.

Segunda etapa — reforçando a negligência. Na segunda etapa, a imagem gerada intermediária é transformada ainda mais em direção ao destino (Figura 25G.2). As ações de edição (Figura 25P.2) reduzem a visibilidade das fachadas de fundo e introduzem árvores e trepadeiras crescidas em excesso, reforçando a impressão de desordem e má conservação. As máscaras de mudança semântica para a etapa final (Figura 25M.V e M.C) destacam as regiões afetadas nas categorias de vegetação e construção, confirmando que as mudanças são localizadas e consistentes com as ações de edição declaradas.

Análise do resultado. A visão de *Estatísticas da Imagem* (Figura 25.IS) confirma a direção perceptual pretendida: em comparação com a origem, a imagem gerada final mostra maior proeminência de construção, vegetação reduzida e desordenada e menor openness, com diminuições correspondentes em imageability, greenness e openness. Aplicar o modelo de pontuação à imagem gerada final produz um Q-score de 3,80, representando uma diminuição de 1,68 ponto em relação à origem e direcionalmente consistente com o escore-alvo de 2,10.

Este cenário demonstra que o valor analítico do UrbanCounterViz não se limita a gerar melhorias. A capacidade de explorar transformações rumo a uma menor segurança percebida apoia a análise diagnóstica e reforça o fio da teoria das Janelas Quebradas que percorre a tese: o sistema pode reproduzir, em forma fotorrealista, os tipos de deterioração ambiental que a teoria e a pesquisa anterior há muito associam a percepções reduzidas de segurança.

7.2 Feedback de Especialistas

Para complementar os cenários de uso com uma perspectiva externa sobre a utilidade e as limitações do sistema, entrevistas semiestruturadas foram conduzidas com três especialistas de domínio que não haviam sido envolvidos no projeto ou no desenvolvimento do UrbanCounterViz. O objetivo era reunir feedback qualitativo sobre três dimensões: usabilidade (se o sistema é claro e compreensível), utilidade (se ele apoia uma análise urbana significativa) e interpretabilidade (se as saídas podem ser interpretadas de forma responsável

para fins de projeto). As entrevistas com especialistas priorizam a profundidade em vez da amplitude — buscando um julgamento fundamentado no domínio sobre o realismo das edições, a relevância prática das recomendações e as implicações éticas de implantar o sistema em contextos de planejamento.

7.2.1 Protocolo de Entrevista

Três especialistas de domínio foram recrutados com experiência em planejamento urbano e análise urbana. O primeiro é um arquiteto com experiência de pesquisa em percepção de segurança em ambientes públicos; o segundo é um pesquisador acadêmico especializado em desenvolvimento urbano contemporâneo; e o terceiro é um especialista em urbanismo. Juntos, esses perfis fornecem perspectivas complementares que abrangem as realidades práticas do projeto e da intervenção urbana, as dimensões analíticas e metodológicas da pesquisa em percepção urbana e o contexto disciplinar mais amplo do urbanismo como campo.

Três entrevistas semiestruturadas separadas, de 45 minutos, foram conduzidas, uma com cada especialista. Cada sessão seguiu a mesma estrutura: uma breve introdução ao processo de coleta de dados e ao pipeline metodológico subjacente ao sistema; uma apresentação detalhada da interface por meio de exemplos representativos e saídas geradas dos cenários de uso; e uma discussão aberta na qual os participantes foram solicitados a refletir sobre a usabilidade, a utilidade, a interpretabilidade e as possíveis limitações do sistema. Os participantes também foram convidados a compartilhar sugestões de melhoria e preocupações sobre como as saídas do sistema poderiam ser interpretadas ou mal utilizadas. Para evitar viés, nenhum dos especialistas é coautor deste trabalho, e nenhum participou do processo de projeto.

7.2.2 Achados

O que se segue sintetiza os principais temas que emergiram ao longo de ambas as entrevistas, organizados pelas três dimensões avaliativas.

Usabilidade.

Todos os três especialistas descreveram o sistema geral como claro e, em geral, intuitivo. As visões *Seletor de Origem* e *Seletor de Destino* foram compreendidas rapidamente, e a *Visão de Caminho* foi apreciada por tornar a sequência de transformação explícita, em vez de apresentar apenas uma comparação de antes e depois. O primeiro especialista observou que as descrições textuais na *Visão de Geração* deveriam refletir mais de perto as modificações reais aplicadas à cena: em alguns casos, as ações de edição declaradas descreviam uma intervenção mais ampla ou ligeiramente diferente do que foi visualmente realizado na imagem gerada, o que poderia enganar analistas que se apoiam no texto

sem examinar a imagem com cuidado. As máscaras de mudança semântica na *Visão de Imagem* foram consideradas úteis, embora o mesmo especialista tenha observado que elas podem ser menos necessárias quando as mudanças visuais já são grandes e imediatamente aparentes, e mais valiosas quando as diferenças são sutis e espacialmente localizadas. O segundo especialista descreveu de forma semelhante o pipeline geral como compreensível e destacou o realismo percebido das imagens geradas como um fator importante para construir confiança nas saídas do sistema. O terceiro especialista levantou uma preocupação sobre usuários iniciantes que abordam a interface sem contexto suficiente, e recomendou que o sistema apresente uma etapa de *onboarding* na qual exemplos representativos de imagens seguras e inseguras sejam mostrados juntamente com breves descrições, permitindo que os usuários calibrem sua compreensão da escala de percepção antes de iniciar qualquer análise.

Utilidade. Todos os três especialistas consideraram o sistema útil para analisar como mudanças visuais localizadas podem afetar a segurança percebida, embora tenham oferecido avaliações matizadas sobre quais tipos de edições são mais praticamente relevantes. O primeiro especialista avaliou positivamente o uso de orientação contextual por meio do grafo de similaridade e dos pares discordantes, observando que fundamentar as edições em diferenças observadas entre cenas reais melhora sua plausibilidade em comparação com a geração sem contexto. O segundo especialista observou que algumas das recomendações geradas poderiam plausivelmente informar intervenções urbanas práticas — particularmente aquelas envolvendo vegetação, pistas de conservação e a remoção de cabos aéreos — enquanto outras, como a reestruturação de vias ou a inserção de grandes elementos construídos, seriam difíceis ou custosas de implementar diretamente e talvez sejam mais bem entendidas como sinais direcionais, e não como propostas acionáveis. Todos os três especialistas enfatizaram que as edições mais convincentes e úteis são aquelas que permanecem sutis e visualmente coerentes dentro da cena; adições proeminentes, como grandes novos edifícios ou veículos em primeiro plano, foram percebidas como menos eficazes, pois podem distorcer a percepção geral da cena e desviar a atenção das mudanças localizadas que são as mais analiticamente significativas. O terceiro especialista adicionalmente destacou o painel de indicadores urbanos das *Estatísticas da Imagem* como um componente particularmente valioso, observando que dimensões como greenness, walkability, enclosure e openness fornecem uma maneira significativa e interpretável de caracterizar as qualidades físicas de um ambiente de rua — uma que vai além do escore de percepção isoladamente e conecta as saídas do sistema a aspectos mensuráveis do espaço urbano sobre os quais os profissionais já raciocinam em contextos de projeto e planejamento.

Considerações éticas. Dois dos três especialistas levantaram preocupações sobre como as saídas do sistema poderiam ser enquadradas e interpretadas na prática. A primeira recomendação é esclarecer o significado dos valores de Q-score exibidos no mapa do *Seletor de Origem*, para evitar que os usuários os interpretem como medições objetivas de segurança

real, em vez de como estimativas de segurança percebida derivadas de julgamentos pareados obtidos por crowdsourcing. Essa distinção — entre segurança percebida e segurança objetiva — nem sempre é imediatamente aparente para usuários não familiarizados com a metodologia do Place Pulse, e confundir as duas poderia levar a interpretações errôneas com implicações reais de política. O terceiro especialista reforçou essa preocupação a partir de um ângulo diferente, recomendando que o sistema contextualize explicitamente os usuários antes de começarem a interagir com a interface — por exemplo, apresentando um conjunto curado de imagens de exemplo que abranjam a faixa de escores de segurança, acompanhadas de breves descrições das características visuais e semânticas que as distinguem. Tal etapa de *onboarding* reduziria o risco de os usuários formarem modelos mentais incorretos sobre o que os escores representam, e ajudaria a evitar danos não intencionais nos casos em que as saídas do sistema são interpretadas como avaliações de perigo real ou objetivo em locais específicos. De forma mais ampla, o feedback dos três especialistas sugere que o valor prático do sistema depende não apenas da qualidade visual e do realismo das edições geradas, mas também de fornecer explicação contextual suficiente para que os analistas possam interpretar as saídas e suas implicações de forma responsável.

7.2.3 Resumo

De modo geral, as entrevistas com especialistas indicam que o UrbanCounterViz é claro, crível e útil para examinar como mudanças visuais localizadas podem influenciar a segurança percebida em ambientes urbanos. O achado mais consistente ao longo das três sessões é que o valor do sistema é maior quando as edições são curtas, visualmente sutis e fundamentadas em diferenças composicionais entre cenas reais — condições que correspondem precisamente às escolhas de projeto enfatizadas ao longo desta tese. Ao mesmo tempo, o feedback identifica duas direções importantes para melhoria: fortalecer o alinhamento entre as ações de edição declaradas e os resultados visuais gerados, e fornecer um enquadramento mais claro das saídas do sistema para evitar a interpretação errônea dos escores de segurança percebida como medições de segurança objetiva.

7.3 Estudo com Usuários

Para complementar as entrevistas com especialistas com uma avaliação mais ampla e estruturada, um estudo com usuários foi conduzido com participantes provenientes de uma gama mais ampla de formações. Enquanto o feedback dos especialistas na Seção 7.2 fornece um julgamento qualitativo fundamentado no domínio, o estudo com usuários visa avaliar como o sistema é percebido pelos tipos de analistas, pesquisadores e profissionais que o encontrariam em um contexto aplicado, e gerar evidência quantitativa de usabilidade que possa ser comparada a referências estabelecidas.

7.3.1 Desenho do Estudo

Um estudo com usuários com 22 participantes de diversas formações acadêmicas e profissionais avaliou a eficácia analítica e a utilidade orientada a valor do URBANCOUNTERVIZ. O estudo recrutou intencionalmente indivíduos sem expertise formal em computação urbana para avaliar se o sistema torna dados complexos de comportamento visual acessíveis a um público mais amplo. Por razões de privacidade, nenhuma informação demográfica foi coletada dos participantes: nem a nacionalidade nem a familiaridade prévia com São Paulo — a cidade da qual as cenas avaliadas são extraídas — foram registradas. Esta é uma escolha de projeto deliberada, mas também uma limitação: como se sabe que a segurança percebida varia com a origem cultural e a familiaridade local (Ceccato, 2012; Ceccato and Loukaitou-Sideris, 2022), não podemos analisar se, por exemplo, os participantes familiarizados com São Paulo julgaram as transformações de forma diferente daqueles que viram as cenas sem contexto local. Retornamos a esse ponto ao discutir a generalizabilidade da avaliação no Capítulo 8. O desenho da avaliação e o questionário foram inspirados no arcabouço ICE-T Wall et al. (2018), que avalia sistemas de visualização em quatro dimensões: Insight, Confidence, Essence e Time. Trabalhos anteriores usando o arcabouço ICE-T sugerem que amostras relativamente pequenas de participantes podem fornecer feedback significativo para avaliações exploratórias de sistemas de visualização Wall et al. (2018), apoiando o uso desse tamanho de amostra. Embora os participantes tivessem experiência prévia em avaliar sistemas interativos, nenhum havia usado o URBANCOUNTERVIZ anteriormente.

7.3.2 Tarefas e Procedimento

O estudo foi conduzido tanto presencialmente quanto online. Os participantes primeiro receberam uma introdução de 10 minutos à interface do URBANCOUNTERVIZ, cobrindo o propósito e as interações de todos os seis componentes visuais: *Seletor de Origem*, *Seletor de Destino*, *Visão de Caminho*, *Visão de Imagem*, *Visão de Geração* e *Estatísticas da Imagem*. Durante essa introdução, os participantes também foram guiados pela seleção de uma imagem de origem e de destino, estabelecendo uma compreensão de base sobre a similaridade entre cenas e as diferenças de escore de segurança (T1.1) antes de prosseguir para as tarefas estruturadas. Os participantes então completaram quatro tarefas de análise visual (U1–U4), projetadas para operacionalizar as tarefas analíticas de alto nível (T1–T3) definidas na Seção 6.3.2.

U1: Identificar similaridades entre cenas e diferenças de composição de imagem exigia que os participantes primeiro usassem o *Seletor de Origem* e o *Seletor de Destino* para identificar e comparar cenas geolocalizadas visualmente similares com base em seus atributos perceptuais e composição semântica (**T1.1**), e então usassem a *Visão de Caminho* para analisar a sequência de transição entre as imagens selecionadas e identificar diferenças na composição de objetos ao longo do caminho (**T1.2**).

U2: Identificar quais componentes mais mudam exigia que os participantes usassem a *Visão de Imagem*, a *Visão de Caminho* e as *Estatísticas da Imagem* para inspecionar e quantificar as diferenças semânticas entre a imagem de origem selecionada e seu contrafactual gerado, operacionalizando a **T2.1**, e para inspecionar como essas diferenças se acumulam ao longo das etapas intermediárias de transição (**T2.2**).

U3: Analisar mudanças específicas e mínimas envolveu inspecionar a composição localizada da imagem por meio das máscaras de segmentação e das descrições textuais na *Visão de Geração*, abordando a **T3.1** ao vincular as ações de edição semântica às regiões destacadas correspondentes na *Visão de Imagem* e ao relacionar as modificações observadas a deslocamentos nos escores perceptuais.

U4: Comparar características visuais das imagens original e gerada exigia que os participantes explorassem o processo completo de geração de múltiplas etapas usando a *Visão de Geração* e as *Estatísticas da Imagem*. Essa tarefa focou em relacionar indicadores urbanos, distribuições de categorias semânticas, escores perceptuais e padrões de similaridade entre as imagens original e gerada (**T3.2**), ao mesmo tempo em que revisitava as diferenças semânticas ao longo das etapas de geração (**T2.2**) para avaliar se a transformação cumulativa se alinhava à direção perceptual pretendida.

7.3.3 Medidas

Para cada tarefa, os participantes foram incentivados a pensar em voz alta Lewis (1982), explicando seu raciocínio e observações antes de responder aos questionários quantitativo (QT) e qualitativo (QL). Além dos itens do ICE-T, os participantes foram solicitados a julgar a mudança de percepção entre cada imagem original e seu contraparte modificada — se a segurança percebida aumentou, diminuiu ou permaneceu inalterada — de modo que sua avaliação da direção da edição pudesse ser comparada à transformação pretendida pelo modelo. O formulário completo usado durante o estudo com usuários, incluindo as quatro tarefas (U1–U4) e essas questões de mudança de percepção, é fornecido no apêndice. Como as tarefas enfatizavam a exploração aberta, o raciocínio visual e a geração de hipóteses, métricas de desempenho tradicionais, como tempo de conclusão da tarefa, taxa de conclusão e frequência de erros, foram intencionalmente omitidas, uma vez que essas medidas não capturam adequadamente a síntese qualitativa e a validação de percepções centrais para os arcabouços de avaliação orientados a valor. Tanto os resultados quantitativos quanto os qualitativos são reportados abaixo.

Questões Quantitativas e Resultados

As questões quantitativas foram diretamente inspiradas no arcabouço ICE-T Wall et al. (2018), cobrindo quatro dimensões: Insight, Confidence, Essence e Time. Cada

dimensão foi avaliada por meio de duas afirmações avaliadas em uma escala Likert de 7 pontos (de "*Totally Disagree*" a "*Totally Agree*").

Insight. "*The combination of the Source Selector and Destination Selector panels made it easier to identify visually similar streets with different safety scores.*" (QT1); "*The Path View and Image Stats panel facilitated the analysis of visual changes between the source and destination images.*" (QT2).

Confidence. "*Using the Image View, Image Stats, and Path View panels: the consistency between the image composition values and the information displayed in the tooltips allowed me to interpret the differences in image composition without ambiguity.*" (QT3); "*Using the Image View, Image Stats, and Generation View panels: the synchronization between the visual differences in the objects, the image composition values, and the textual explanations allowed me to easily verify the changes made to the source image.*" (QT4).

Essence. "*The tool's set of visualizations made it possible to quickly obtain an overview of the visual characteristics and differences among the analyzed images.*" (QT5); "*The combination of the visualization of image changes (Image View) and textual explanations (Generation View) made it possible to quickly understand the key alterations made to the images.*" (QT6).

Time. "*Using the Image View, Image Stats, and Generation View panels, the tool's visualizations and textual descriptions enabled quick identification of which elements underwent the most significant changes between the original and generated images.*" (QT7); "*The image selection options (Source Selector and Destination Selector), the visualization of composition differences (tooltip), and the navigation between panels allowed me to identify similar images and analyze their differences quickly and efficiently.*" (QT8).

A Tabela 12 apresenta os escores ICE-T por participante ao longo de todas as quatro dimensões, onde cada escore representa a média dos dois itens Likert correspondentes. De modo geral, as respostas concentram-se predominantemente na extremidade positiva da escala, com feedback neutro ou negativo mínimo. As médias das dimensões — Insight: 6,11, Confidence: 6,48, Essence: 6,23, Time: 6,36 — todas excedem 6,0 na escala de 7 pontos, indicando avaliações consistentemente favoráveis em todas as quatro dimensões. A dimensão Confidence recebeu o maior escore médio, sugerindo que os participantes acharam a coordenação entre os componentes visuais, os valores de composição e as descrições textuais particularmente eficaz para interpretar e verificar as saídas do sistema. A dimensão Insight recebeu a média comparativamente mais baixa, com escores individuais variando de forma mais ampla do que nas outras dimensões (de 5,0 a 7,0), apontando para alguma variabilidade em quão facilmente os participantes identificaram cenas visualmente similares com escores de segurança diferentes durante o estágio inicial de seleção. Essa variabilidade alinha-se às sugestões de melhoria da interface reportadas nos achados qualitativos abaixo.

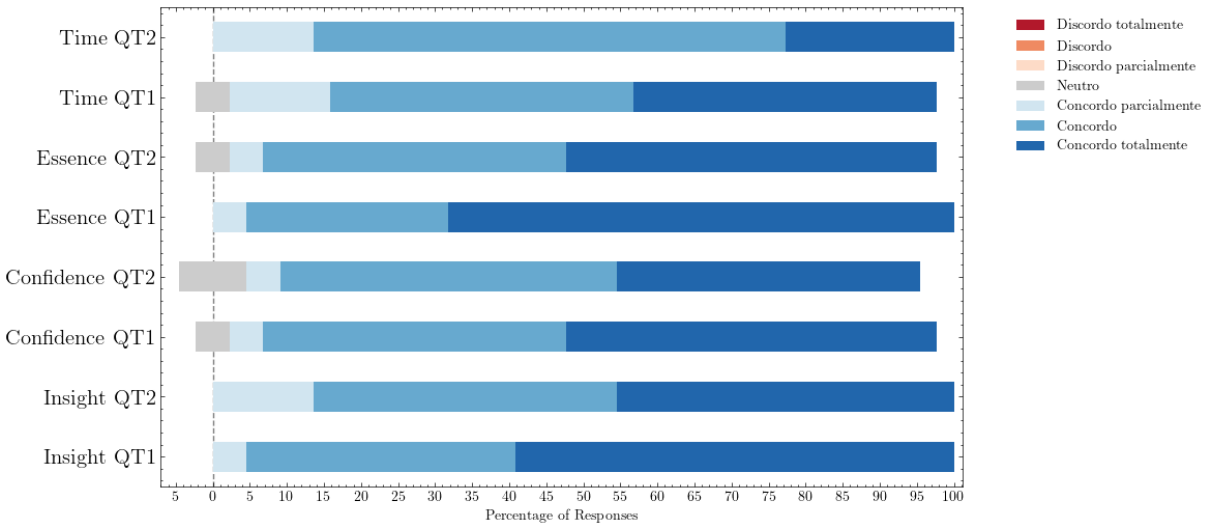

Figure 26 – Distribuição das respostas dos participantes para as questões QT1–QT8, agrupadas por dimensões inspiradas no ICE-T (Insight, Confidence, Essence e Time). O comprimento da barra indica a porcentagem de participantes que selecionaram cada nível Likert em uma escala de 7 pontos, de "*Totally Disagree*" a "*Totally Agree*".

Table 12 – Escores da avaliação ICE-T por participante ao longo das quatro dimensões. Cada escore é a média de dois itens Likert (1–7).

Participant	***Insight***	***Confidence***	***Essence***	***Time***
P1	6.5	7.0	6.0	7.0
P2	7.0	7.0	6.5	6.0
P3	6.5	6.5	5.5	7.0
P4	6.0	5.5	5.5	6.0
P5	5.5	6.0	7.0	6.5
P6	6.0	7.0	6.5	6.5
P7	5.0	7.0	6.0	7.0
P8	6.5	7.0	7.0	7.0
P9	6.5	6.0	7.0	5.5
P10	6.0	6.5	7.0	6.0
P11	6.0	6.5	5.5	5.0
P12	5.5	6.5	6.0	6.5
P13	5.5	5.0	5.0	5.0
P14	6.0	7.0	7.0	6.5
P15	6.5	7.0	4.0	6.5
P16	6.0	7.0	7.0	6.0
P17	6.0	6.0	6.0	6.0
P18	7.0	6.5	7.0	6.0
P19	6.5	6.5	6.0	7.0
P20	6.0	7.0	6.0	7.0
P21	5.5	6.0	6.5	7.0
P22	6.5	6.0	7.0	7.0
Avg.	**6.11**	**6.48**	**6.23**	**6.36**

A Figura 26 confirma ainda mais a distribuição das respostas ao longo da escala Likert.

Questões Qualitativas e Achados

O questionário qualitativo coletou feedback aberto sobre a utilidade do UrbanCounterViz e oportunidades de melhorar a interface por meio de duas questões:

"*Considering the system's main objective — analyzing the improvement and worsening of perception of a street — which feature, graph, or interaction was most useful for your analyses, and why?*" (QL1); "*What would you change, add, or remove from the interface to make it more efficient? Feel free to include any other comments or suggestions regarding your experience.*" (QL2).

As respostas às QL1 e QL2 identificaram consistentemente a *Visão de Imagem* e a *Visão de Geração* como os componentes mais úteis para comparar imagens, identificar diferenças visuais e compreender as descrições semânticas associadas a cada etapa de geração. Os participantes também enfatizaram que os recursos interativos, particularmente os tooltips, ajudaram a reduzir a poluição visual, ao mesmo tempo em que preservavam o acesso à informação contextual.

Quanto aos painéis de dados, os participantes relataram que o painel de composição semântica das *Estatísticas da Imagem* era intuitivo para compreender as mudanças nos objetos detectados, enquanto o painel de indicadores urbanos exigia maior esforço cognitivo porque sua interpretação dependia de valores numéricos, em vez de pistas visuais diretas. Esse achado sugere que as representações visuais de objetos são mais imediatamente interpretáveis do que os resumos quantitativos de indicadores, e aponta para uma possível direção de redesenho do painel de indicadores urbanos — por exemplo, por meio de codificações baseadas em ícones ou em cores que reduzam a dependência da leitura numérica.

As sugestões de melhoria focaram principalmente em aumentar o controle do usuário sobre o processo de edição. Em vez de apenas inspecionar as ações e descrições textuais geradas pelo sistema, os participantes expressaram interesse em propor e avaliar suas próprias modificações interativamente. Recomendações adicionais incluíram permitir que os painéis fossem redimensionados e aumentar a proeminência da visão de mapa para facilitar a seleção inicial dos locais de origem e destino, consistente com os escores de Insight mais baixos observados nos resultados quantitativos.

7.3.4 Resumo

O estudo com usuários fornece evidência quantitativa e qualitativa estruturada de que o UrbanCounterViz é eficaz e acessível a um público amplo de analistas e pesquisadores sem expertise especializada em computação urbana. Os escores ICE-T em todas as quatro dimensões excederam 6,0 em uma escala de 7 pontos — uma faixa que indica avaliações positivas, mas não acríticas, consistente com participantes que identificam tanto

pontos fortes quanto áreas concretas para melhoria. Esses escores refletem um sistema que é eficaz e acessível, ao mesmo tempo em que retém espaço para refinamento, particularmente no fluxo de trabalho de seleção de cenas e no painel de indicadores urbanos. A dimensão Confidence recebeu o maior escore médio (6,48), corroborando o achado do feedback dos especialistas de que a coordenação entre os componentes visuais e as descrições textuais é o ativo de usabilidade mais forte do sistema. A *Visão de Imagem* e a *Visão de Geração* foram as mais frequentemente citadas como os componentes mais valiosos, reforçando a escolha de projeto de tornar o processo de edição explícito e inspecionável a cada etapa, em vez de apresentar apenas um resultado final.

Ao mesmo tempo, o estudo identifica direções claras para melhoria. A dimensão Insight recebeu o escore médio comparativamente mais baixo e a maior dispersão de respostas individuais, com sugestões qualitativas sobre a proeminência do mapa e a redimensionabilidade dos painéis apontando para atrito no estágio inicial de seleção de cenas — o ponto de entrada do fluxo de trabalho analítico. O painel de indicadores urbanos, embora apreciado em princípio pelos especialistas de domínio, exigiu mais esforço cognitivo do que o painel de composição semântica entre a amostra mais ampla de participantes, sugerindo que sua codificação visual poderia ser tornada mais imediata. Por fim, o pedido recorrente por controle de edição interativo — a capacidade de propor e testar as próprias modificações, em vez de apenas inspecionar as geradas pelo sistema — alinha-se diretamente à direção de trabalho futuro identificada no Capítulo 8: possibilitar ações de edição direcionáveis pelo usuário como um próximo passo de desenvolvimento para o UrbanCounterViz.

7.4 Considerações Finais

Este capítulo apresentou a avaliação do UrbanCounterViz por meio de três métodos complementares: cenários de uso, feedback de especialistas e um estudo com usuários. Tomados em conjunto, eles avaliam o sistema a partir de ângulos diferentes, porém mutuamente reforçadores — demonstrando seu alcance analítico por meio de cenários estruturados, validando sua credibilidade prática por meio do julgamento de especialistas de domínio e medindo sua usabilidade e cobertura de requisitos por meio de uma amostra mais ampla de participantes.

Vários achados merecem ênfase particular, pois conectam a avaliação de volta às contribuições da tese como um todo. Primeiro, a estratégia de fundamentação — derivar edições de diferenças entre cenas adjacentes reais, em vez de geração irrestrita — foi consistentemente identificada, ao longo dos três métodos de avaliação, como a propriedade mais responsável pela interpretabilidade e credibilidade do sistema. Segundo, a transparência explícita do pipeline — a capacidade de inspecionar diferenças semânticas, ações de edição, imagens geradas e resumos multidimensionais de resultados a cada etapa

— foi reconhecida como essencial para que o fluxo de trabalho com humano no processo funcione como pretendido. Terceiro, a capacidade bidirecional demonstrada no Cenário 3 estende o sistema para além do planejamento de melhorias, rumo à análise diagnóstica, possibilitando uma compreensão mais completa dos fatores visuais que moldam a percepção de segurança em ambas as direções. Quarto, o estudo com usuários fornece fundamentação quantitativa para as observações qualitativas das entrevistas com especialistas, fornecendo evidência preliminar sobre se a usabilidade e a cobertura de requisitos percebidas pelos profissionais de domínio também se sustentam para uma população de usuários mais ampla.

O Capítulo 8 sintetiza esses achados com as contribuições mais amplas da tese. Ele situa os resultados da avaliação dentro do arcabouço de quatro contribuições estabelecido no Capítulo 1 — a revisão sistemática, a análise de elementos visuais, o pipeline UrbanCFX (Capítulo 5) e o sistema UrbanCounterViz com seu fluxo de trabalho interpretável reutilizável — e reflete sobre até que ponto cada contribuição aborda as lacunas metodológicas identificadas no Capítulo 3. O Capítulo 9 então tira as conclusões finais, reconhece as limitações do trabalho como um todo e delineia direções para pesquisas futuras.

8 Discussão

As quatro contribuições desta tese — a revisão sistemática da literatura (Capítulo 3), a investigação dos elementos visuais na percepção de segurança por meio do UrbanFormer (Capítulo 4), o pipeline UrbanCFX e o conjunto de dados UrbanPD4k (Capítulo 5) e o sistema UrbanCounterViz com seu fluxo de trabalho interpretável reutilizável (Capítulos 6–7) — abordam, cada uma, um aspecto distinto, porém relacionado, do desafio da análise interpretável da percepção de segurança urbana. Tomadas individualmente, cada uma faz uma contribuição significativa: a revisão sistemática mapeia o panorama metodológico do campo e identifica as lacunas que esta tese aborda; o UrbanFormer identifica quais elementos visuais impulsionam a atenção do modelo durante a predição de segurança; o UrbanCFX traduz contrafactuais do espaço de características em recomendações em linguagem natural; e o UrbanCounterViz gera transformações de cenas urbanas fotorrealistas, fundamentadas e interativamente exploráveis. No entanto, as percepções mais importantes emergem ao considerar essas contribuições em conjunto, como estágios em um argumento de pesquisa cumulativo sobre o que significa compreender, explicar e agir sobre estimativas computacionais da percepção de segurança urbana.

Este capítulo sintetiza essas percepções. A Seção 8.1 integra os principais achados de todas as contribuições e identifica as lições abrangentes que elas oferecem para o campo da pesquisa em percepção urbana. A Seção 8.2 discute as limitações do pipeline de pesquisa completo, indo além das limitações em nível de capítulo já apresentadas em cada contribuição. A Seção 8.3 delineia direções de pesquisa futuras que decorrem diretamente do trabalho aqui apresentado.

8.1 Síntese dos Achados

8.1.1 Fundamentando o Raciocínio Contrafactual em Dados Urbanos Reais

Um fio que percorre as três contribuições empíricas é o princípio de que a análise interpretável da percepção urbana deve ser fundamentada em diferenças observadas entre cenas reais, em vez de derivada da otimização do modelo ou da geração irrestrita. Esse princípio foi estabelecido teoricamente na Seção 6.4.3 por meio do conceito formal de pares discordantes — enquanto a intuição por trás de comparar cenas similares com escores divergentes já motivava as análises dos Capítulos 4 e 5 — e realizado como um sistema totalmente interativo e guiado por caminho no UrbanCounterViz.

Os resultados da avaliação no Capítulo 7 fornecem evidência concreta de por

que a fundamentação importa. Nos cenários de uso, as edições derivadas de pares discordantes reais produziram transformações mais visualmente coerentes e analiticamente interpretáveis do que as trajetórias mais longas, nas quais a restrição de fundamentação enfraquecia à medida que o comprimento do caminho aumentava. Os especialistas de domínio corroboraram essa observação, identificando independentemente a fundamentação contextual como a propriedade mais responsável pela credibilidade prática do sistema. Essa convergência entre a análise quantitativa dos cenários e o julgamento qualitativo dos especialistas fortalece o argumento de que a estratégia de fundamentação não é meramente uma escolha metodológica, mas um contribuinte substantivo para o valor do sistema.

De uma perspectiva mais ampla, esse achado conecta-se a uma limitação bem documentada dos métodos de explicação contrafactual em aprendizado de máquina: contrafactuais irrestritos frequentemente produzem entradas que são ótimas para deslocar a predição de um modelo, mas irrealistas ou ininterpretáveis como intervenções do mundo real (Wachter et al., 2017; Mothilal et al., 2020). A abordagem proposta nesta tese aborda essa limitação no domínio da percepção urbana ao substituir contrafactuais sintéticos por contrafactuais fundamentados em observação, usando a estrutura de um grafo de similaridade visual para restringir as transformações ao espaço de mudanças que de fato ocorrem entre cenas urbanas reais.

8.1.2 O Papel da Representação Semântica

Ao longo das três contribuições empíricas, a representação semântica da cena desempenhou um papel fundamental — mas em diferentes níveis de granularidade e para diferentes propósitos. No Capítulo 4, o vetor de razão de pixels de 150 categorias do ADE20K possibilitou a identificação de quais categorias amplas de objetos (árvores, calçadas, muros, solo exposto) estavam mais associadas a predições seguras e inseguras. No Capítulo 5, esse vocabulário foi estendido com 13 elementos de desordem anotados manualmente (pichação, cabos aéreos, calçadas quebradas, muros danificados) que são invisíveis aos modelos de segmentação padrão, mas altamente relevantes para os julgamentos humanos de segurança. No UrbanCounterViz (Capítulos 6–7), os dois vocabulários são combinados: as categorias do ADE20K fornecem a espinha dorsal composicional para a construção do grafo de similaridade, a guia por caminho e o cálculo de diferenças semânticas, enquanto o vocabulário de elementos de desordem do UrbanPD4k informa o estágio de geração de ações de edição e enriquece os resumos semânticos exibidos na interface.

Esse enriquecimento progressivo do vocabulário semântico reflete uma percepção analítica genuína: as pistas visuais que mais fortemente diferenciam ambientes urbanos seguros de inseguros são frequentemente distinções *intracategoria* que as taxonomias de segmentação padrão confundem. Um muro conservado e um muro coberto de pichação ocupam a mesma categoria do ADE20K, mas produzem julgamentos de segurança muito

diferentes. Uma calçada ladeada por árvores e uma trilha de terra nua ambas se registram como combinações de *floor* e *vegetation*, mas sinalizam níveis muito diferentes de ordem ambiental. O esforço de anotação de desordem no Capítulo 5 e sua integração ao UrbanCounterViz representam uma resposta metodológica a essa limitação, e os cenários de uso no Capítulo 7 demonstram a consequência: as edições mais informativas e analiticamente críveis nos Cenários 1 e 3 envolvem precisamente essas transformações intracategoria — a introdução de um muro deteriorado coberto de pichação, a substituição de cercas vivas organizadas por vegetação crescida em excesso.

8.1.3 Transparência e Supervisão Humana como Propriedades do Sistema

Um dos achados consistentes ao longo da avaliação é que a transparência — a capacidade de compreender não apenas o que o sistema produz, mas por quê — não é um recurso periférico, mas um determinante central da utilidade do sistema. Os especialistas de domínio no Capítulo 7 não avaliaram o UrbanCounterViz primariamente pela qualidade visual de suas imagens geradas, mas por se conseguiam compreender e confiar no raciocínio por trás delas. Essa observação tem implicações importantes para o projeto de sistemas analíticos em domínios de alto risco.

Os sistemas de análise visual anteriores para imagens urbanas, como o StreetVizor (Shen et al., 2017) e o Urban Mosaic (Miranda et al., 2020), priorizaram a exploração e a recuperação em detrimento da explicação. Os métodos generativos anteriores, como o FaceLift (Joglekar et al., 2020) e o URSimulator (Hu et al., 2025), priorizaram a qualidade visual em detrimento da interpretabilidade da lógica da transformação. O UrbanCounterViz é projetado desde o início para tornar cada etapa de seu pipeline inspecionável: a estrutura do grafo de similaridade, a seleção de pares discordantes, o cálculo de diferenças semânticas, a geração de ações de edição, a saída fotorrealista e o resumo multidimensional de resultados são todos expostos ao analista por meio de visões coordenadas. Os resultados da avaliação sugerem que essa transparência não é meramente uma virtude de projeto, mas um pré-requisito funcional para que o sistema seja considerado crível e útil pelos profissionais de domínio.

O sistema também reforça um ponto mais amplo sobre a relação entre a análise visual e a interpretabilidade do aprendizado de máquina. O arcabouço VIS4ML (Sacha et al., 2019) e trabalhos relacionados sobre IA explicável (Choo and Liu, 2018; La Rosa et al., 2023) argumentaram que a interação com humano no processo é essencial para a implantação responsável de modelos de ML em domínios consequentes. O projeto e a avaliação do UrbanCounterViz fornecem um estudo de caso concreto no qual esse princípio é aplicado ao domínio da percepção urbana: o analista humano não é um consumidor passivo das saídas do modelo, mas um participante ativo que seleciona caminhos, interpreta etapas intermediárias e decide quando uma transformação gerada é

analiticamente crível.

8.1.4 Bidirecionalidade como Ferramenta Diagnóstica

O Cenário 3 no Capítulo 7 demonstrou uma dimensão do valor analítico do UrbanCounterViz que não foi explicitamente antecipada nos requisitos de projeto, mas que emergiu naturalmente da capacidade bidirecional do sistema: a capacidade de gerar transformações rumo a uma *menor* segurança percebida como meio de diagnosticar quais pistas visuais mais fortemente sinalizam desordem. Essa direção de análise é complementar aos cenários orientados à melhoria e conecta o sistema a uma longa tradição na pesquisa urbana de compreender a deterioração e a desordem como processos ativos, e não meramente como a ausência de ordem.

De uma perspectiva da teoria das Janelas Quebradas (Wilson and Kelling, 1982), que motiva a tese desde seu capítulo inicial, a capacidade de gerar e inspecionar cenários de deterioração é tão analiticamente importante quanto a capacidade de gerar melhorias. Compreender por que uma cena é percebida como insegura exige compreender não apenas quais elementos estão associados à segurança (árvores, calçadas conservadas, ambiente construído organizado), mas também quais elementos sinalizam ativamente a desordem (pichação, crescimento excessivo, superfícies deterioradas). O Cenário 3 mostra que o UrbanCounterViz pode gerar esses sinais de desordem em forma fotorrealista, fundamentados em pares discordantes reais, e rastrear suas fontes semânticas por meio das visões coordenadas da interface. Isso dá aos pesquisadores uma ferramenta para explorar e formular hipóteses sobre a desordem visual — por exemplo, se a pichação isoladamente parece suficiente para deslocar uma cena de segura para insegura em uma transformação gerada, ou se ela tende a exigir a coocorrência de desordem da vegetação e deterioração estrutural.

8.2 Limitações

As limitações a seguir aplicam-se ao pipeline de pesquisa completo apresentado nesta tese, estendendo-se para além das limitações em nível de capítulo já discutidas nas Seções 4.4.4 e 5.5, e das limitações de avaliação identificadas na Seção 7.4.

Generalizabilidade dos achados. Todas as contribuições empíricas apoiam-se principalmente no Place Pulse 2.0, um conjunto de dados coletado por meio de *crowdsourcing* online a partir de uma população global, mas demograficamente não representativa. As percepções de segurança capturadas neste conjunto de dados refletem os julgamentos agregados de uma população de anotadores específica em um momento específico, e podem não se generalizar entre contextos culturais, grupos demográficos ou cidades com diferentes características visuais e de infraestrutura. A análise do UrbanCFX no Capítulo 5 restringe

ainda mais seu escopo ao Rio de Janeiro, cujas características socioespaciais — particularmente a coocorrência de assentamentos informais, elementos de desordem visíveis e fortes contrastes na qualidade da infraestrutura — podem não ser representativas de outros contextos urbanos. Os cenários de uso no Capítulo 7 usam São Paulo, que compartilha algumas dessas características, mas a validação sistemática intercidades permanece uma questão em aberto importante.

Percepção versus segurança objetiva. Ao longo desta tese, o termo "*safety perception*" é usado de forma consistente e cuidadosa para se referir a julgamentos perceptuais humanos derivados de informação visual, e não a taxas de criminalidade reais, riscos físicos ou outras medidas objetivas de segurança. No entanto, como o feedback dos especialistas na Seção 7.2 destacou, essa distinção é facilmente borrada na prática, particularmente quando os escores perceptuais são exibidos em um mapa que superficialmente se assemelha a um mapa de criminalidade ou de risco. Futuras implantações do UrbanCounterViz deveriam investir na comunicação, em nível de interface, dessa distinção — por exemplo, por meio de rotulagem explícita, tooltips contextuais e materiais de *onboarding* — para garantir que as saídas sejam interpretadas de forma responsável por planejadores e formuladores de políticas que podem não estar familiarizados com a metodologia de percepção obtida por crowdsourcing.

Latência de geração. Cada etapa de edição no pipeline VLM de dois estágios requer aproximadamente 18 segundos para ser concluída, principalmente devido ao tempo de inferência do modelo de edição de imagem Gemini. Para trajetórias curtas (duas a três etapas), isso é gerenciável dentro de uma única sessão analítica. Para trajetórias mais longas (cinco a seis etapas), a latência acumulada se aproxima de dois minutos, o que interrompe o fluxo analítico e limita o número de caminhos alternativos que um analista pode explorar em um período de tempo razoável. Essa restrição é particularmente relevante para o caso de uso com humano no processo que o UrbanCounterViz é projetado para apoiar: se o custo de explorar uma única trajetória alternativa é de vários minutos de espera, os analistas podem convergir prematuramente para a primeira trajetória explorada, em vez de comparar múltiplas opções.

Dependência de modelos de geração proprietários. Uma limitação relacionada diz respeito à dependência de um modelo de geração proprietário e de código fechado. O pipeline de edição de dois estágios usa modelos Gemini acessados por meio de uma API comercial, o que restringe a reprodutibilidade: as versões do modelo podem mudar ou ser descontinuadas, a inferência não é totalmente determinística e o processo de geração interno não pode ser auditado independentemente. Embora a estratégia de fundamentação e o pipeline circundante sejam, por projeto, agnósticos ao modelo, replicar os resultados visuais específicos reportados nesta tese exige acesso à mesma versão do modelo usada durante a avaliação.

Controle espacial na edição de imagem. O modelo VLM de edição de imagem exibe precisão espacial limitada sobre as regiões semânticas especificadas. Em várias instâncias ao longo dos cenários de uso — particularmente em trajetórias mais longas — o modelo introduziu grandes elementos em primeiro plano ou reestruturou globalmente a geometria da cena de maneiras não implicadas pelo vetor de diferença semântica. As máscaras baseadas em segmentação ajudam os analistas a identificar essas mudanças não intencionais após o fato, mas não as previnem durante a geração. Trabalhos futuros integrando *inpainting* condicionado espacialmente ou modelos de difusão guiados por máscara poderiam abordar essa limitação ao restringir as edições a regiões específicas da imagem, em vez de permitir que o modelo generativo modifique a cena globalmente.

Escalabilidade da anotação. O esforço de anotação de desordem de 13 categorias no Capítulo 5 exigiu dois anotadores treinados — ambos estudantes de doutorado com experiência prévia em percepção urbana e residência de longo prazo no Rio de Janeiro — trabalhando em 3.659 imagens, com reconciliação para resolver desacordos. Embora esse esforço tenha enriquecido substancialmente o vocabulário semântico disponível para o sistema, ele não é diretamente escalável para coleções de imagens maiores ou para novas cidades onde os elementos de desordem relevantes podem diferir. Modelos automatizados de detecção de desordem treinados no conjunto de dados UrbanPD4k poderiam abordar esse gargalo, mas exigiriam uma validação cuidadosa para garantir que as características específicas do domínio — como a aparência visual da habitação informal em São Paulo versus Rio de Janeiro versus outras cidades — sejam adequadamente capturadas.

Imagens estáticas e a negligência da variação temporal e climática. Uma limitação que atravessa todas as contribuições desta tese é que todas as análises se baseiam em instantâneos de *street view* únicos e estáticos. Cada localização é representada por uma imagem, capturada em um momento, sob um conjunto de condições de clima e iluminação. Isso tem uma consequência direta e facilmente ignorada para os resultados: no Capítulo 4, o céu emergiu como uma categoria de alta prevalência, mas quase neutra, e seria tentador concluir que o céu simplesmente não importa para a segurança percebida. Essa conclusão seria um artefato da representação, e não um achado sobre a percepção. Nossas características codificam apenas *quanto* céu é visível, nunca seu *estado* — nublado versus limpo, luz do dia intensa versus crepúsculo, seco versus chuva — porque esses fatores são mantidos fixos pelo projeto de instantâneo único e descartados pelo vetor de razão de pixels. A criminologia ambiental mostrou precisamente o oposto da neutralidade para esses fatores: a segurança percebida, o medo do crime e até a distribuição espaço-temporal do crime violento variam sistematicamente com o clima, a estação e a hora do dia, incluindo evidências de São Paulo de que os padrões de homicídio acompanham a variação climática e temporal, e trabalhos mais amplos que estabelecem como o ritmo temporal do ambiente molda o medo e a percepção do espaço urbano (Ceccato, 2005; Wikström et al., 2010; Ceccato and Uittenbogaard, 2014; Ceccato and Loukaitou-Sideris,

2022). A aparente neutralidade do céu nesta tese deve, portanto, ser lida de forma restrita, como uma afirmação sobre a *área* de céu visível em um conjunto de dados de condição fixa, e não como uma alegação de que as condições do céu são perceptualmente irrelevantes. Capturar esses efeitos é uma direção importante para trabalhos futuros e exigiria imagens amostradas temporalmente (por exemplo, a mesma localização em diferentes horas do dia, estações e condições climáticas) juntamente com características sensíveis à aparência, em vez de descritores baseados apenas em presença.

Relevância ponderada por presença e o subcrédito de elementos pequenos. A medida de relevância em nível de elemento introduzida no Capítulo 4 — a sobreposição entre a atenção do modelo e cada máscara de segmentação, calculada em média ao longo do conjunto de dados — é inerentemente *ponderada por presença*, e isso merece mais ênfase do que seu tratamento em nível de capítulo lhe confere. Como a relevância é quantificada como sobreposição de área espacial e, então, calculada em média ao longo de todas as imagens, categorias grandes e ubíquas, como árvores ou edifícios, acumulam escores altos em parte em virtude da mera área de imagem que ocupam, enquanto um elemento pequeno que cobre apenas cerca de 1% de uma cena — pichação, uma placa quebrada, um saco de lixo — é empurrado para um impacto médio próximo de zero, mesmo nas imagens onde o modelo atende a ele fortemente. A métrica, em outras palavras, recompensa *quanto* da imagem um elemento preenche, em vez de quão *saliente* ele é onde aparece, o que arrisca subcreditar sistematicamente exatamente as pistas de desordem compactas que o enquadramento das Janelas Quebradas coloca no centro da segurança percebida. Isso não é meramente um incômodo de escala: significa que o ranqueamento de relevância não deve ser lido como um ranqueamento direto da importância perceptual para objetos pequenos. Abordar isso exigiria desacoplar a frequência de ocorrência da saliência por objeto — por exemplo, normalizando a sobreposição pela própria área de cada elemento, condicionando a média apenas às imagens onde o elemento está presente, ou complementando a IoU baseada em área com medidas de densidade de atenção que recompensem a ativação concentrada sobre uma pequena região. Consideramos isso um refinamento metodológico concreto para trabalhos de explicação futuros, e não um aspecto consolidado do pipeline atual.

Pichação como categoria única: desordem versus arte. Ao longo do vocabulário de desordem desta tese, a pichação é tratada como uma categoria e, implicitamente, como um sinal de negligência. Isso justifica uma qualificação mais forte do que o capítulo de anotação sozinho fornece. A pichação não é uma pista homogênea: a marcação (*tagging*) ilícita e o vandalismo situam-se em uma extremidade, mas o muralismo, a arte pública encomendada e a arte de rua sancionada situam-se na outra, e estas últimas são frequentemente percebidas não como desordem alguma, mas como evidência de investimento, identidade cultural e um espaço público ativo e cuidado — e podem, portanto, aumentar, em vez de diminuir, a segurança percebida. Como o presente trabalho colapsa esses casos em uma única classe

anotada, e como os estágios contrafactual e de edição herdam essa classe, o sistema pode tratar a remoção de qualquer pichação como uma melhoria inequívoca, quando em alguns contextos o oposto poderia se sustentar. Isso é uma limitação conceitual genuína da taxonomia de desordem, e não apenas uma questão de custo de anotação: um tratamento mais fiel subcategorizaria a pichação por tipo (por exemplo, *tags* versus murais) e estudaria se a pichação artística se comporta como um elemento distinto, potencialmente aumentador da segurança. Deixamos esse refinamento do vocabulário de desordem para trabalhos futuros.

8.3 Trabalhos Futuros

As limitações identificadas acima sugerem várias direções concretas para pesquisas futuras, organizadas das extensões técnicas mais imediatas às direções metodológicas e de aplicação mais amplas.

Ações de edição direcionáveis pelo usuário. A extensão mais imediata do UrbanCounterViz é permitir que os analistas refinem ou substituam as ações de edição geradas no Estágio 1 do pipeline antes que sejam passadas ao modelo de edição de imagem. Atualmente, a *Visão de Geração* exibe as ações para inspeção, mas o analista não pode modificá-las. Uma interface tipo chat na qual os analistas pudessem ajustar a redação do prompt, excluir categorias específicas ou modular a magnitude das mudanças propostas transformaria a geração em um processo mais genuinamente colaborativo e reduziria a lacuna entre as edições pretendidas e as realizadas, identificada no feedback dos especialistas.

Manipulação contrafactual de forma livre. Para além de direcionar o pipeline existente, trabalhos futuros poderiam permitir que os analistas especifiquem as composições semânticas desejadas diretamente — por exemplo, ajustando controles deslizantes na visão de *Estatísticas da Imagem* que representem proporções-alvo para cada categoria semântica — em vez de navegar apenas por um grafo de similaridade pré-computado. Essa capacidade expandiria o espaço de transformações exploráveis preservando o enquadramento analítico do sistema, e complementaria a abordagem guiada por caminho ao possibilitar a exploração orientada por hipóteses, além da descoberta orientada por dados.

Adaptação intercidades e ao contexto local. Como observado no Capítulo 5, a restrição a uma única cidade na análise do UrbanCFX foi uma escolha metodológica deliberada para eliminar confundidores intercidades e garantir uma investigação controlada dos elementos de desordem. Estender essa abordagem a múltiplas cidades permanece, ainda assim, uma direção em aberto importante. O sistema atual é instanciado no Place Pulse 2.0, que fornece ampla cobertura geográfica, mas especificidade local limitada. Estender o sistema para incorporar conjuntos de dados de percepção específicos de cidade ou distrito

— coletados por meio de pesquisas com moradores, plataformas de *crowdsourcing* locais ou estudos móveis — fundamentaria suas recomendações nos julgamentos de segurança das comunidades mais diretamente afetadas pelos ambientes urbanos analisados. Essa direção é particularmente importante para garantir que as saídas do UrbanCounterViz sejam responsivas a gramáticas visuais culturalmente específicas e a pistas de desordem localmente significativas, em vez de refletirem os julgamentos agregados de um conjunto de anotadores globalmente distribuído, mas demograficamente não representativo.

Integração com fluxos de trabalho de projeto urbano. Os cenários de uso no Capítulo 7 e o feedback dos especialistas na Seção 7.2 sugerem que as saídas do UrbanCounterViz — particularmente as ações de edição e as imagens geradas a partir de trajetórias curtas e coerentes — poderiam funcionar como entradas para processos de projeto urbano, em vez de resultados analíticos isolados. Trabalhos futuros poderiam explorar como as recomendações do sistema poderiam ser exportadas em formatos compatíveis com ferramentas de projeto urbano (por exemplo, visões em nível de rua anotadas, relatórios de mudança semântica ou listas de intervenção priorizadas), e poderiam avaliar se as intervenções informadas pelo UrbanCounterViz melhoram a segurança percebida em estudos de campo longitudinais.

Avaliação com amostras mais amplas de participantes. O estudo de feedback com especialistas no Capítulo 7 envolveu três participantes, o que é apropriado para uma investigação qualitativa baseada em entrevistas, mas não fornece o poder estatístico necessário para generalizar conclusões sobre usabilidade ou utilidade. Trabalhos futuros deveriam incluir estudos com usuários mais extensos, com participantes que abranjam uma gama mais ampla de expertise de domínio, de planejadores urbanos e arquitetos profissionais a moradores da comunidade e funcionários do governo local. Isso forneceria um quadro mais completo de como diferentes populações de usuários interpretam as saídas do UrbanCounterViz e agem sobre elas.

8.4 Considerações Finais

A alegação central desta tese é que compreender a percepção de segurança urbana exige mais do que modelos preditivos acurados: exige ferramentas que apoiem o raciocínio interpretável, fundamentado e orientado ao planejamento sobre os fatores visuais que impulsionam os julgamentos humanos dos ambientes urbanos. A evidência reunida ao longo dos Capítulos 3 a 7 — desde as lacunas evidenciadas na literatura, aos elementos identificados pelo UrbanFormer (Capítulo 4), às recomendações em linguagem simples do UrbanCFX (Capítulo 5), às transformações inspecionáveis produzidas pelo UrbanCounterViz (Capítulos 6–7) — fala a favor dessa alegação, ainda que uma única dissertação só possa levá-la até certo ponto.

A síntese apresentada neste capítulo sugere três lições mais amplas para o campo. Primeiro, fundamentar o raciocínio contrafactual em dados reais — por meio do arcabouço de pares discordantes e da estrutura do grafo de similaridade — não é meramente uma preferência metodológica, mas um contribuinte substantivo para a interpretabilidade e a credibilidade das recomendações resultantes. Segundo, o vocabulário semântico importa: as distinções intracategoria que as taxonomias de segmentação padrão deixam passar (muros conservados versus deteriorados, vegetação organizada versus crescida em excesso) são precisamente as pistas que mais fortemente diferenciam cenas urbanas seguras de inseguras na percepção humana. Terceiro, a transparência e a supervisão humana não são recursos periféricos dos sistemas analíticos para domínios de alto risco, mas determinantes centrais de sua utilidade, como demonstrado pela ênfase consistente na inspecionabilidade do pipeline tanto nos cenários de uso quanto no feedback dos especialistas.

O Capítulo 9 conclui a tese revisitando os quatro objetivos específicos estabelecidos no Capítulo 1, resumindo as contribuições em relação a cada objetivo e refletindo sobre as implicações mais amplas deste trabalho para o estudo da percepção urbana, o projeto de sistemas de IA interpretáveis para o ambiente construído e o papel da análise visual em conectar modelos computacionais com o conhecimento de domínio orientado ao planejamento.

9 Conclusões

Esta dissertação propôs-se a compreender o que faz uma rua parecer segura ou insegura, e a perguntar se essa compreensão poderia ser transformada em algo mais útil do que uma predição. Os capítulos que se seguiram perseguiram essa questão em três passos, delineados pela revisão sistemática que abriu o trabalho: movendo-se da identificação de elementos visuais, à descrição de mudanças plausíveis a eles em linguagem simples, até, por fim, tornar essas mudanças visíveis e inspecionáveis. O que a evidência reunida ao longo do caminho sustenta é que isso é possível, e que cada passo é genuinamente necessário: nenhum dos três, isoladamente, teria sido suficiente.

9.1 Contribuições Revisitadas

Cada um dos quatro objetivos específicos estabelecidos no Capítulo 1 — mapear o estado da pesquisa em percepção urbana, identificar os elementos visuais que moldam a segurança percebida, traduzir essas associações em recomendações legíveis por humanos e tornar as mudanças plausíveis visíveis e inspecionáveis — é atendido pela contribuição correspondente revisitada a seguir.

A revisão sistemática no Capítulo 3 triou 2.307 artigos e sintetizou 234 estudos, mapeando o panorama metodológico da pesquisa em percepção urbana utilizando SVI e identificando quatro lacunas persistentes — percepção limitada sobre os fatores visuais, explicações estáticas, edições generativas irrealistas e a ausência de ferramentas interativas com humano no processo — que motivaram diretamente todas as contribuições subsequentes.

O Capítulo 4 abordou a primeira lacuna por meio do UrbanFormer, que combinou representações de CNN com vetores de razão de pixels do OneFormer e aplicou a análise Grad-CAM em mais de 108.000 imagens para identificar quais elementos visuais — árvores, calçadas, muros, solo exposto — mais fortemente moldam as predições de segurança percebida. O Capítulo 5 aprofundou isso com o UrbanCFX e o conjunto de dados UrbanPD4k, introduzindo 13 elementos de desordem física anotados manualmente e um pipeline que traduz vetores de diferença contrafactual em recomendações de projeto em linguagem simples por meio de um LLM, fechando a lacuna comunicativa entre as saídas do modelo e os profissionais urbanos.

Os Capítulos 6 e 7 introduziram o UrbanCounterViz, um sistema de análise visual que estende o raciocínio contrafactual para o domínio da imagem. Ao fundamentar cada edição em diferenças observadas entre pares de cenas discordantes reais, derivar transformações fotorrealistas por meio de um pipeline VLM de dois estágios e expor todo o

processo por meio de uma interface coordenada de seis visões, o UrbanCounterViz torna possível uma exploração de cenas urbanas fundamentada, inspecionável e direcionável. A avaliação por meio de cenários de uso, entrevistas com especialistas e um estudo com 22 participantes forneceu evidência convergente do valor analítico do sistema, com escores ICE-T excedendo 6,0 em todas as quatro dimensões e especialistas de domínio identificando consistentemente a fundamentação e a transparência como as propriedades mais responsáveis pela credibilidade do sistema.

Ao lado do próprio UrbanCounterViz, a quarta contribuição desta tese também inclui um fluxo de trabalho interpretável reutilizável — independente de qualquer implementação de sistema específica — que formaliza como as predições do modelo, as diferenças semânticas de cena e as transformações visuais fundamentadas podem ser sistematicamente conectadas. Esse fluxo de trabalho, ancorado no conceito de par discordante e no grafo de similaridade visual, fornece um modelo generalizável para trabalhos futuros sobre a análise interpretável da percepção urbana.

9.2 Considerações Finais

Os ambientes urbanos moldam o comportamento e o bem-estar de maneiras consequentes. A capacidade de raciocinar computacionalmente sobre quais características visuais sinalizam segurança percebida ou desordem — e de explorar como mudanças plausíveis nessas características poderiam deslocar o julgamento humano — não é, portanto, meramente uma conquista técnica, mas uma contribuição prática para o projeto urbano e a política pública.

Tomadas em conjunto, essas contribuições reenquadram a análise da segurança urbana percebida como algo mais do que uma tarefa de modelagem preditiva. Elas mostram como modelos de percepção orientados a dados, representações semânticas de cena, raciocínio contrafactual, edição generativa de imagens e análise visual podem ser conectados em um único fluxo de trabalho para o raciocínio fundamentado, interpretável e inspecionável por humanos sobre os ambientes visuais urbanos.

O arcabouço estabelecido nesta tese — fundamentar a análise contrafactual em diferenças reais observadas, tornar cada estágio do pipeline transparente e inspecionável, e conectar o comportamento do modelo a recomendações orientadas ao planejamento — fornece uma base fundamentada para extensões futuras: generalização intercidades, edição direcionável pelo usuário e integração com fluxos de trabalho de projeto profissionais. Essas permanecem como direções em aberto; a base sobre a qual elas podem ser construídas é a contribuição deste trabalho.

REFERENCES

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning From Data*. AMLBook.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Acosta, S. F. and Camargo, J. E. (2018). City safety perception model based on visual content of street images. *2018 IEEE International Smart Cities Conference (ISC2)*, pages 1–8.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Alain, G. and Bengio, Y. (2018). Understanding intermediate layers using linear classifier probes, 2018. *URL https://arxiv. org/abs/1610.01644*, 1610.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Alzate, J. R., Tabares, M. S., and Vallejo, P. (2021). Graffiti and government in smart cities: a deep learning approach applied to medellin city, colombia. In *International Conference on Data Science, E-learning and Information Systems 2021*, pages 160–165.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*.

Andersson, V. O., Birck, M. A., and Araujo, R. M. (2017). Investigating crime rate prediction using street-level images and siamese convolutional neural networks. In *Latin American Workshop on Computational Neuroscience*, pages 81–93. Springer.

Arietta, S. M., Efros, A. A., Ramamoorthi, R., and Agrawala, M. (2014). City forensics: Using visual elements to predict non-visual city attributes. *IEEE transactions on visualization and computer graphics*, 20(12):2624–2633.

Augustin, M., Boreiko, V., Croce, F., and Hein, M. (2022). Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.

Beaucamp, B., Leduc, T., Tourre, V., and Servieres, M. C. J. (2022). The whole is other than the sum of its parts: Sensibility analysis of 360° urban image splitting. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Beneduce, C., Lepri, B., and Luca, M. (2025). Urban safety perception through the lens of large multimodal models: A persona-based approach. *Miniworkshop: Mobile and Social Computing Lab*.

Bengio, Y. (2007). Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2:1–127.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

Ceccato, V. (2005). Homicide in sao paulo, brazil: Assessing spatial-temporal and weather variations. *Journal of Environmental Psychology*, 25(3):307–321.

Ceccato, V. (2012). The urban fabric of crime and fear. In *The urban fabric of crime and fear*, pages 1–33. Springer.

Ceccato, V. and Loukaitou-Sideris, A. (2022). Fear of sexual harassment and its impact on safety perceptions in transit environments: a global perspective. *Violence against women*, 28(1):26–48.

Ceccato, V. and Uittenbogaard, A. C. (2014). Space–time dynamics of crime in transport nodes. *Annals of the Association of American geographers*, 104(1):131–150.

Charreire, H., Mackenbach, J. D., Ouasti, M., Lakerveld, J., Compernolle, S., Ben-Rebah, M., McKee, M., Brug, J., Rutter, H., and Oppert, J.-M. (2014). Using remote sensing to define environmental characteristics related to physical activity and dietary behaviours: a systematic review (the spotlight project). *Health & place*, 25:1–9.

Chen, L.-C. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Choo, J. and Liu, S. (2018). Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications*, 38(4):84–92.

Collaris, D. and van Wijk, J. J. (2020). ExplainExplore: Visual exploration of machine learning explanations. In *Proc. PacificVis*, pages 26–35.

Costa, G., Soares, C., and Marques, M. (2019). Finding common image semantics for urban perceived safety based on pairwise comparisons. *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31.

Diniz, A. M. A. and Stafford, M. C. (2021). Graffiti and crime in belo horizonte, brazil: The broken promises of broken windows theory. *Applied Geography*, 131:102459.

Doersch, C., Singh, S., Gupta, A. K., Sivic, J., and Efros, A. A. (2012). What makes paris look like paris? *ACM Transactions on Graphics (TOG)*, 31:1 – 9.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dubey, A., Naik, N., Parikh, D., Raskar, R., and Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 196–212, Cham. Springer International Publishing.

Ebtekar, A. and Liu, P. (2021). Elo-mmr: A rating system for massive multiplayer competitions. In *Proceedings of the Web Conference 2021*, pages 1772–1784.

Gomez, O., Holter, S., Yuan, J., and Bertini, E. (2020). ViCE: visual counterfactual explanations for machine learning models. In *Proc. IUI*, IUI '20, pages 531–535, New York, NY, USA. Association for Computing Machinery.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Google (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR.

Han, S. and Kim, D. (2025). Multimodal geoai for urban perception prediction: A case study of songdo, incheon. *Journal of Digital Landscape Architecture*, pages 645–652.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, N. and Li, G. (2021). Urban neighbourhood environment assessment based on street view image processing: A review of research trends. *Environmental Challenges*, 4:100090.

Hinkle, J. C. and Weisburd, D. (2008). The irony of broken windows policing: A micro-place study of the relationship between disorder, focused police crackdowns and fear of crime. *Journal of Criminal justice*, 36(6):503–512.

Hu, C., Jia, S., and Li, X. (2025). URSimulator: Human-perception-driven prompt tuning for enhanced virtual urban renewal via diffusion models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 228:356–369.

Hu, C., Jia, S., Zhang, F., Xiao, C., Ruan, M., Thrasher, J., and Li, X. (2023). UPDExplainer: An interpretable transformer-based framework for urban physical disorder detection using street view imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:209–222.

Huang, W., Wang, J., and Cong, G. (2024). Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *International Journal of Geographical Information Science*, 38:1414 – 1442.

Ibrahim, M. R., Haworth, J., and Cheng, T. (2020). Understanding cities with machine eyes: A review of deep computer vision in urban analytics. *Cities*, 96:102481.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. Random House, New York.

Jain, J., Li, J., Chiu, M. T., Hassani, A., Orlov, N., and Shi, H. (2023). Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998.

Jeanneret, G., Simon, L., and Jurie, F. (2022). Diffusion models for counterfactual explanations. In *Proc. ACCV*, pages 858–876.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Joglekar, S., Quercia, D., Redi, M., Aiello, L. M., Kauer, T., and Sastry, N. R. (2020). Facelift: a transparent deep learning framework to beautify urban scenes. *Royal Society Open Science*, 7.

Keizer, K., Lindenberg, S., and Steg, L. (2008). The spreading of disorder. *Science (New York, N.Y.)*, 322:1681–5.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. (2024). Debating with more persuasive llms leads to more truthful answers. In *41st International Conference on Machine Learning*.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR.

La Rosa, B., Blasilli, G., Bourqui, R., Auber, D., Santucci, G., Capobianco, R., Bertini, E., Giot, R., and Angelini, M. (2023). State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum*, 42(1):319–355.

Lavi., B., Tokuda., E., Moreno-Vera., F., Nonato., L., Silva., C., and Poco., J. (2022). 17k-graffiti: Spatial and crime data assessments in são paulo city. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,*, pages 968–975. INSTICC, SciTePress.

Law, S., Hasegawa, R., Paige, B., Russell, C., and Elliott, A. (2023). Explaining holistic image regressors and classifiers in urban analytics with plausible counterfactuals. *International Journal of Geographical Information Science*, 37(12):2575–2596.

Law, S., Paige, B., and Russell, C. (2018a). Take a look around. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10:1 – 19.

Law, S., Seresinhe, C. I., Shen, Y., and Gutierrez-Roig, M. (2018b). Street-frontage-net: urban image classification using deep convolutional neural networks. *International Journal of Geographical Information Science*, 34:681 – 707.

Levering, A., Marcos, D., Jacobs, N., and Tuia, D. (2024). Prompt-guided and multimodal landscape scenicness assessments with vision-language models. *PLOS ONE*, 19.

Lewis, C. (1982). Using the "thinking aloud"method in cognitive interface design. *Research Report RC9265, IBM TJ Watson Research Center*.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022a). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Li, X., Zhang, C., and Li, W. (2015a). Does the visibility of greenery increase perceived safety in urban areas? evidence from the place pulse 1.0 dataset. *ISPRS Int. J. Geo Inf.*, 4:1166–1183.

Li, X., Zhang, C., Li, W., Ricard, R. M., Meng, Q., and Zhang, W. (2015b). Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening*, 14:675–685.

Li, Y., Peng, L., chong Wu, C., and Zhang, J. (2022b). Street view imagery (svi) in the built environment: A theoretical and systematic review. *Buildings*.

Li, Y., Zhong, X., Liu, H., Liao, M., Lu, Y.-f., and Liu, B. (2025). Urban built-up area extraction using triangle threshold algorithm and naive bayes classification model with multidata fusion. *Scientific Reports*, 15(1):40175.

Li, Z., Liu, P., Shi, J., and Xing, Y. (2021). Research on street space quality combined with attention multi-task deep learning. *2021 2nd International Conference on Big Data Economy and Information Management (BDEIM)*, pages 434–441.

Liang, X., Zhao, T., and Biljecki, F. (2023). Revealing spatio-temporal evolution of urban visual environments with street view imagery. *Landscape and Urban Planning*.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Lindal, P. J. and Hartig, T. (2013). Architectural variation, building height, and the restorative quality of urban residential streetscapes. *Journal of Environmental Psychology*, 33:26–36.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Liu, C., Jang, K. M., Dimitrov, S., Barbalho, C. L. F., Duarte, F., and Ratti, C. (2025). Unveiling the internal structures of informal settlements through ai-driven automated segmentation of lidar data. *npj Urban Sustainability*, 5(1):108.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023a). Visual instruction tuning. In *NeurIPS*.

Liu, W., Guo, W., and Wu, R.-X. (2022a). Understanding urban wealth perception using a hybrid dataset and ranking-scoring framework. *Transactions in GIS*, 26:2366 – 2382.

Liu, Y., Chen, M., Wang, M., Huang, J., Thomas, F., Rahimi, K., and Mamouei, M. (2023b). An interpretable machine learning framework for measuring urban perceptions from panoramic street view images. *iScience*, 26.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022b). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Neural Information Processing Systems*.

Lynch, K. (1984). Reconsidering the image of the city. *Cities of the Mind*, pages 151–161.

Ma, H., Li, J., and Ye, X. (2025). Deep learning meets urban design: Assessing streetscape aesthetic and design quality through AI and cluster analysis. *Cities*, 162:105939.

Ma, H. and Wu, D. (2023). A natural language processing-based approach: mapping human perception by understanding deep semantic features in street view images. *ArXiv*, abs/2311.17354.

Ma, X., Ma, C., Wu, C., Xi, Y., Yang, R., Peng, N., Zhang, C., and Ren, F. (2021). Measuring human perceptions of streetscapes to better inform urban renewal: A perspective of scene semantic parsing. *Cities*, 110:103086.

Ma, Z. (2023). Deep exploration of street view features for identifying urban vitality: A case study of qingdao city. *Int. J. Appl. Earth Obs. Geoinformation*, 123:103476.

Malekzadeh, M. S., Willberg, E. S., Torkko, J., and Toivonen, T. (2024). Urban visual appeal according to chatgpt: Contrasting ai and human insights. *ArXiv*, abs/2407.14268.

Marasinghe, R., Yigitcanlar, T., Mayere, S., Washington, T., and Limb, M. (2023). Computer vision applications for urban planning: A systematic review of opportunities and constraints. *Sustainable Cities and Society*, page 105047.

Min, W., Mei, S., Liu, L., Wang, Y., and Jiang, S. (2020). Multi-task deep relative attribute learning for visual urban perception. *IEEE Transactions on Image Processing*, 29:657–669.

Minka, T., Cleven, R., and Zaykov, Y. (2018). Trueskill 2: An improved bayesian skill rating system. Technical Report MSR-TR-2018-8, Microsoft.

Miranda, F., Hosseini, M., Lage, M., Doraiswamy, H., Dove, G., and Silva, C. T. (2020). Urban mosaic: Visual exploration of streetscapes using large-scale image data. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.

Molnar, C. (2022). *Interpretable Machine Learning*. Lulu.com, 2 edition.

Moreno-Vera, F., Brandoli, B., and Poco, J. (2024). What makes a place feel safe? analyzing street view images to identify relevant visual elements. *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

Moreno-Vera, F., De-la Puente, A., and Poco, J. (2025). UrbanPhysicalDisorder-4k: Understanding urban perception via counterfactuals and street view signs of physical disorder. In *Proc. BigData*, pages 5194–5200.

Moreno-Vera, F., Lavi, B., and Poco, J. (2021). Quantifying urban safety perception on street view images. In *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

Moreno-Vera, F. and Poco, J. (2025). Assessing urban environments with vision-language models: A comparative analysis of AI-generated ratings and human volunteer evaluations. In *Proc. IJCNN*, pages 1–8.

Moreno-Vera, F. and Poco, J. (2026). From vision-centric to multimodal: A systematic review of street view–based urban perception. In *IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI)*.

Mothilal, R. K., Sharma, A., and Tan, C. (2019). Explaining machine learning classifiers through diverse counterfactual explanations. *2020 Conference on Fairness, Accountability, and Transparency*.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617.

Muller, E., Gemmell, E., Choudhury, I., Nathvani, R. S., Metzler, A. B., Bennett, J. E., Denton, E. L., Flaxman, S., and Ezzati, M. (2022). City-wide perceptions of neighbourhood quality using street view images. *ArXiv*, abs/2211.12139.

Naik, N., Philipoom, J., Raskar, R., and Hidalgo, C. (2014). StreetScore: predicting the perceived safety of one million streetscapes. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Naik, N., Raskar, R., and Hidalgo, C. A. (2016). Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *The American Economic Review*, 106:128–132.

Nasar, J. L. (1998). The evaluative image of the city. *Journal of the American Planning Association*.

Nasar, J. L. and Jones, K. M. (1997). Landscapes of fear and stress. *Environment and behavior*, 29(3):291–323.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ordonez, V. and Berg, T. L. (2014). Learning high-level judgments of urban perception. *European Conference on Computer Vision (ECCV)*.

Pan, C., Li, H., Wang, L., Wu, J., Guo, J., Qiu, N., and Liu, X. (2024). Data science basis and influencing factors for the evaluation of environmental safety perception in macau parishes. *Advances in Continuous and Discrete Models*, 2024(1):44.

Park, J. and Newman, M. (2005). A network-based ranking system for us college football. *Journal of Statistical Mechanics: Theory and Experiment*, 2005:P10014 – P10014.

Porzi, L., Rota Bulò, S., Lepri, B., and Ricci, E. (2015). Predicting and understanding urban perception with convolutional neural networks. *MM '15: Proceedings of the 23rd ACM international conference on Multimedia*.

Qiu, W., Li, W., Liu, X., Zhang, Z., Li, X., and Huang, X. (2023). Subjective and objective measures of streetscape perceptions: Relationships with property value in shanghai. *Cities*, 132:104037.

Quercia, D., O'Hare, N., and Cramer, H. (2014). Aesthetic capital: what makes london look beautiful, quiet, and happy? *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Ramírez, S. (2019). Fastapi framework, high performance, easy to learn, fast to code, ready for production.

Regona, M., Yigitcanlar, T., Xia, B., and Li, R. Y. M. (2022). Opportunities and adoption challenges of ai in the construction industry: a prisma review. *Journal of Open Innovation: Technology, Market, and Complexity*, 8(1):45.

Remigio, R. V., Zulaika, G., Rabello, R. S., Bryan, J., Sheehan, D. M., Galea, S., Carvalho, M. S., Rundle, A., and Lovasi, G. S. (2019). A local view of informal urban environments: A mobile phone-based neighborhood audit of street-level factors in a brazilian informal community. *Journal of Urban Health*, 96(4):537–548.

Rzotkiewicz, A., Pearson, A. L., Dougherty, B. V., Shortridge, s., and Wilson, N. (2018). Systematic review of the use of google street view in health research: Major themes, strengths, weaknesses and possibilities for future research. *Health & place*, 52:240–246.

Sacha, D., Kraus, M., Keim, D. A., and Chen, M. (2019). VIS4ML: An ontology for visual analytics assisted machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):385–395.

Salesses, P., Schechtner, K., and Hidalgo, C. A. (2013). The collaborative image of the city: Mapping the inequality of urban perception. *PLOS ONE*.

Sampson, R. J., Morenoff, J. D., and Gannon-Rowley, T. (2002). Assessing "neighborhood effects": Social processes and new directions in research. *Annual review of sociology*, 28(1):443–478.

Sangers, R., van Gemert, J. C., and van Cranenburgh, S. (2022). Explainability of deep learning models for urban space perception. *ArXiv*, abs/2208.13555.

Schroeder, H. W. and Anderson, L. M. (1984). Perception of personal safety in urban recreation sites. *Journal of leisure research*, 16(2):178–194.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Seresinhe, C. I., Preis, T., and Moat, H. S. (2017). Using deep learning to quantify the beauty of outdoor places. *Royal Society Open Science*, 4.

Shen, Q., Zeng, W., Ye, Y., Arisona, S. M., Schubiger, S., Burkhard, R., and Qu, H. (2017). Streetvizor: Visual exploration of human-scale urban forms based on street views. *IEEE transactions on visualization and computer graphics*, 24(1):1004–1013.

Shneiderman, B. (2002). The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. IEEE Symposium on Visual Languages*, pages 364–371. IEEE.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Skogan, W. G. (1992). *Disorder and decline: Crime and the spiral of decay in American neighborhoods*. Univ of California Press.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise.

Song, Q., Fang, Y., Li, M., van Ameijde, J., and Qiu, W. (2025). Exploring the coherence and divergence between the objective and subjective measurement of streetscape perceptions at the neighborhood level: A case study in shanghai. *Environment and Planning B: Urban Analytics and City Science*, 52(5):1231–1251.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Stalidis, P., Semertzidis, T., and Daras., P. (2018). Examining deep learning architectures for crime classification and prediction. *arXiv*.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tan, X., Song, Q., Liu, X., and Qiu, W. (2025). Visual perception-informed urban design toolkit: Computational urban morphology optimisation to inform real-time perceived safety. *Journal of Urban Management*.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Thurstone, L. L. (2017). A law of comparative judgment. In *Scaling*, pages 81–92. Routledge.

Tokuda, E. K., Silva, C. T., and Jr., R. M. C. (2019). Quantifying the presence of graffiti in urban environments. *CoRR*, abs/1904.04336.

Ulrich, R. S. (1979). Visual landscapes and psychological well-being. *Landscape research*, 4(1):17–23.

Vago, N. O. P., Milani, F., Fraternali, P., and da Silva Torres, R. (2021). Comparing cam algorithms for the identification of salient image features in iconography artwork analysis. *Journal of Imaging*, 7.

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *arXiv preprint arXiv:1711.00399*.

Wall, E., Agnihotri, M., Matzen, L., Divis, K., Haass, M., Endert, A., and Stasko, J. (2018). A heuristic approach to value-driven evaluation of visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):491–500.

Wang, Y., Zeng, Z., Li, Q., and Deng, Y. (2022a). A complete reinforcement-learning-based framework for urban-safety perception. *ISPRS Int. J. Geo Inf.*, 11:465.

Wang, Y., Zeng, Z., and Zhao, Q. (2022b). Evaluating the perceived safety of urban city via maximum entropy deep inverse reinforcement learning. In *Asian Conference on Machine Learning*.

Wang, Y., Zhou, L., Wang, Y., and Peng, Z. (2024). Leveraging pretrained language models for enhanced entity matching: A comprehensive study of fine-tuning and prompt learning paradigms. *International Journal of Intelligent Systems*, 2024(1):1941221.

Wikström, P.-O. H., Ceccato, V., Hardie, B., and Treiber, K. (2010). Activity fields and the dynamics of crime: Advancing knowledge about the role of the environment in crime causation. *Journal of quantitative criminology*, 26(1):55–87.

Wilson, J. Q. and Kelling, G. L. (1982). Broken windows. *Atlantic monthly*, 249(3):29–38.

Wilson, J. S., Kelly, C. M., Schootman, M., Baker, E. A., Banerjee, A., Clennin, M. N., and Miller, D. K. (2012). Assessing the built environment using omnidirectional imagery. *American journal of preventive medicine*, 42 2:193–9.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pages 1–10.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wu, C., Liang, Y., Zhao, M., Teng, M., Han, Y., and Ye, Y. (2025). Perceiving the fine-scale urban poverty using street view images through a vision-language model. *Sustainable Cities and Society*, 123(106267).

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090.

Yao, Y., Dall'Ò, G., and Lu, F. (2026). Urban street-scene perception and renewal strategies powered by vision–language models. *Land*, 15(2):244.

Yao, Z. and Zhu, T. (2025). Dynamic urban imagery and emotional perception in shanghai's north bund: A deep learning approach. *Advances in Education, Humanities and Social Science Research*.

Yin, C., Peng, N., Li, Y., Shi, Y., Yang, S., and Jia, P. (2023). A review on street view observations in support of the sustainable development goals. *International Journal of Applied Earth Observation and Geoinformation*, 117:103205.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zhang, C., Wu, T., Zhang, Y., Zhao, B., Wang, T., Cui, C., and Yin, Y. (2021). Deep semantic-aware network for zero-shot visual urban perception. *International Journal of Machine Learning and Cybernetics*, 13:1197 – 1211.

Zhang, F., Hu, M., Che, W., Lin, H., and Fang, C. (2018a). Framework for virtual cognitive experiment in virtual geographic environments. *ISPRS Int. J. Geo Inf.*, 7:36.

Zhang, F., Salazar-Miranda, A., Duarte, F., Vale, L., Hack, G., Chen, M., Liu, Y., Batty, M., and Ratti, C. (2024). Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery. *Annals of the American Association of Geographers*, pages 1–22.

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., and Ratti, C. (2018b). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180:148–160.

Zhang, J., Li, Y., Fukuda, T., and Wang, B. (2025a). Urban safety perception assessments via integrating multimodal large language models with street view images. *Cities*, 165:106122.

Zhang, Y., Xiong, X., Yang, S., Zhang, Q., Chi, M., Wen, X., Zhang, X., and Wang, J. (2025b). Enhancing the visual environment of urban coastal roads through deep learning analysis of street-view images: A perspective of aesthetic and distinctiveness. *PLOS ONE*, 20.

Zhanga, J., Lia, Y., Fukudab, T., and Wang, B. (2024). Revolutionizing urban safety perception assessments: Integrating multimodal large language models with street view images. *ArXiv*, abs/2407.19719.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890.

Zhao, X., Lu, Y., and Lin, G. (2024). An integrated deep learning approach for assessing the visual qualities of built environments utilizing street view images. *Engineering Applications of Artificial Intelligence*, 130:107805.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Zhou, H., He, S., Cai, Y., Wang, M., and Su, S. (2019). Social inequalities in neighborhood visual walkability: Using street view imagery and deep learning technologies to facilitate healthy city planning. *Sustainable Cities and Society*.

Zhou, H., Wang, J., Wilson, K., Widener, M., Wu, D. Y., and Xu, E. (2025a). Using street view imagery and localized crowdsourcing survey to model perceived safety of the visual built environment by gender. *Int. J. Appl. Earth Obs. Geoinformation*, 139:104421.

Zhou, Q., Zhang, J., and Zhu, Z. (2025b). Evaluating urban visual attractiveness perception using multimodal large language model and street view images. *Buildings*.

Zhu, Y., Su, F., Han, X., Fu, Q., and Liu, J. (2024). Uncovering the drivers of gender inequality in perceptions of safety: An interdisciplinary approach combining street view imagery, socio-economic data and spatial statistical modelling. *Int. J. Appl. Earth Obs. Geoinformation*, 134:104230.

Zytek, A., Pido, S., Alnegheimish, S., Berti-Équille, L., and Veeramachaneni, K. (2024). Explingo: Explaining ai predictions using large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1197–1208.