Felipe Adrian Moreno Vera

Exploring Urban Safety Perception with Vision-Language Models and Semantically Grounded Image Counterfactuals

Rio de Janeiro

June, 2026

Felipe Adrian Moreno Vera

Exploring Urban Safety Perception with Vision-Language Models and Semantically Grounded Image Counterfactuals

Thesis submitted to the School of Applied Mathematics (FGV/EMAp) as a requirement for the PhD degree in Applied Mathematics and Data Science.

Research Area: Urban computing, Generative AI, Vision-Language Models (VLMs), and Visual Analytics.

Getulio Vargas Foundation
School of Applied Math

Supervisor PhD. Jorge Poco

Rio de Janeiro
June, 2026

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Vera, Felipe Adrian Moreno
Exploring urban safety perception with vision-language models and semantically grounded image counterfactuals / Felipe Adrian Moreno Vera. –2026.

141 f.

Tese (doutorado) – Escola de Matemática Aplicada Mestrado em Modelagem Matemática da Informação.
Orientador: Jorge Medina
Inclui bibliografia.

1. Planejamento urbano. 2. Inteligência artificial. 3. Análise de imagem. 4. Semântica - Processamento de dados 5. Aprendizado do computador I. Medina, Jorge. II. Escola de Matemática Aplicada. III. Título.

CDD – 006.3

Elaborada por Maria do Socorro Almeida – CRB-7/4254

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA
DOUTORADO EM MATEMÁTICA APLICADA E CIÊNCIA DE DADOS
FOLHA DE APROVAÇÃO

FELIPE ADRIAN MORENO VERA

"EXPLORING URBAN SAFETY PERCEPTION WITH VISION-LANGUAGE MODELS AND SEMANTICALLY GROUNDED IMAGE COUNTERFACTUALS".

TESE APRESENTADA AO CURSO DE DOUTORADO EM MATEMÁTICA APLICADA E CIÊNCIA DE DADOS QUE CONFERIU O GRAU DE DOUTOR EM MATEMÁTICA APLICADA E CIÊNCIA DE DADOS.

DATA DA DEFESA: 27/07/2026

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

PRESIDENTE DA COMISSÃO EXAMINADORA: PROF. JORGE LUIS POCO-MEDINA

<ASSINADO ELETRONICAMENTE>
PROF. JORGE LUIS POCO-MEDINA
ORIENTADOR

<ASSINADO ELETRONICAMENTE>
PROF. DÁRIO AUGUSTO BORGES OLIVEIRA
MEMBRO INTERNO -FGV EMAP

<ASSINADO ELETRONICAMENTE>
PROF. JOÃO LUCAS RULFF DA COSTA
MEMBRO INTERNO - FGV EMAP

<ASSINADO ELETRONICAMENTE>
PROF. RICARDO HURTUBIA
MEMBRO EXTERNO - PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

<ASSINADO ELETRONICAMENTE>
PROF. LUIS GUSTAVO NONATO
MEMBRO EXTERNO - USP

RIO DE JANEIRO, 27 DE JULHO DE 2026.

<ASSINADO ELETRONICAMENTE>
PROF. CÉSAR LEOPOLDO CAMACHO MANCO
DIRETOR

<ASSINADO ELETRONICAMENTE>
PROF FLAVIO CARVALHO DE VASCONCELOS
PRÓ-REITOR

ESTE É UM TRABALHO ORIGINAL ONDE FOI VERIFICADA A NÃO EXISTÊNCIA DE PLÁGIO E DE UTILIZAÇÃO DE INTELIGÊNCIA ARTIFICIAL, NÃO EXPLICITADA, NO CORPO DO TRABALHO ATESTADO PELO ALUNO(A) E ORIENTADOR(A). ESTE DOCUMENTO NÃO CONFERE TÍTULO. PARA TAL DEVERÃO SER CUMPRIDOS OS REQUISITOS DO PROGRAMA DE PÓS-GRADUAÇÃO.

Acknowledgements

This research is part of a Ph.D. program at the Getulio Vargas Foundation (FGV). First, I would like to thank the Inner Circle, also known as the "golden generation," as well as my advisor for their guidance, the Visualization Data Science Lab (Visual-DS) for their brainstorming and support, and my family for their continued support from afar.

Moreover, this work was supported by the Carlos Chagas Filho Foundation for Research Support of the State of Rio de Janeiro (FAPERJ), Brazil; the São Paulo Research Foundation (FAPESP), Brazil; the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES); and the School of Applied Mathematics at Fundação Getulio Vargas (FGV/EMAp −− VisualDS).

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of FAPESP, FAPERJ, CAPES, or FGV.

"To my family,
who encouraged me to achieve my goals.
:)"

Felipe A. Moreno

Statement of Authorship

I hereby declare that the thesis submitted is my original work. All direct or indirect sources used are acknowledged as references. In addition, I declare that I have not submitted this thesis to any other institution to obtain a degree.

—

Acronyms

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

Figure 1 – ML types . . . 25
Figure 2 – Grad-CAM . . . 29
Figure 3 – Perception Collection . . . 34
Figure 4 – Comparisons to perceptual scores . . . 38
Figure 5 – Systematic review flowchart . . . 43
Figure 6 – Keyword wordcloud . . . 46
Figure 7 – Author affiliations vs applications . . . 47
Figure 8 – Sankey diagram of author affiliations . . . 49
Figure 9 – Number of studies per task category . . . 51
Figure 10 – UrbanFormer model . . . 54
Figure 11 – Segmentation model masks . . . 58
Figure 12 – UrbanFormer explanation . . . 61
Figure 13 – IoU technique . . . 62
Figure 14 – UrbanFormer results . . . 65
Figure 15 – UrbanCFX idea . . . 68
Figure 16 – UrbanPD4k dataset . . . 73
Figure 17 – Spatial distribution of perceived safety in Rio de Janeiro . . . 73
Figure 18 – UrbanPD4k statistics . . . 76
Figure 19 – UrbanCounterViz Teaser . . . 82
Figure 20 – UrbanCounterViz workflow . . . 88
Figure 21 – UrbanCounterViz Interface . . . 98
Figure 22 – UrbanCounterViz scenario 1 . . . 104
Figure 23 – UrbanCounterViz scenario 2 . . . 105
Figure 24 – UrbanCounterViz multistep evaluation . . . 107
Figure 25 – UrbanCounterViz scenario 3 . . . 108
Figure 26 – Distribution of participant responses for questions QT1–QT8 . . . 115

List of Tables

Table 1 – Publications and contributions 22
Table 2 – Place Pulse 2.0 statistics 41
Table 3 – Number of papers 44
Table 4 – SVI providers 50
Table 5 – Learning tasks 52
Table 6 – UrbanFormer binary results 63
Table 7 – UrbanFormer 10-label results 63
Table 8 – ADE20K classes and new group-categories (visual elements) 71
Table 9 – Classification results using binary and pixel ratios configurations 78
Table 10 – LLM human-readable interpretations for Counterfactuals 79
Table 11 – UrbanCounterViz Analytical Tasks 97
Table 12 – UrbanCounterViz Quantitative questions results 116

List of Prompts

Prompt 1 UrbanCFX prompt template 75
Prompt 2 UrbanCounterViz action prompt template 94
Prompt 3 UrbanCounterViz image editing prompt template 96

Contents

Acknowledgements 4

Statement of Authorship 5

Acronyms 6

Abstract 7

Resumo 8

List of Figures 9

List of Tables 10

List of Prompts 11

1 INTRODUCTION 16
1.1 Thesis statement 19
1.2 Objectives 19
1.3 Contributions 20
1.4 Outline 22

2 BACKGROUND 24
2.1 Machine Learning and Deep Learning 24
2.2 Interpretability and Explainability 28
2.2.1 Explanation Methods 28
2.2.2 Interpretability and Counterfactual Methods 30
2.3 Multimodal Learning and Vision-Language Models (VLMs) 32
2.3.1 Multimodal Representation Learning 32
2.3.2 Vision-Language Models (VLMs) 32
2.3.3 Generative and Instruction-Following VLMs 33
2.4 Urban Perception and AI 34
2.4.1 Perception collection 35
2.4.2 Perception calculation 35
2.4.3 Perception tasks 39
2.4.4 Perception dataset 40
2.5 Final Considerations 41

3 SYSTEMATIC LITERATURE REVIEW 42

3.1 Methodology . . . 42
3.1.1 Identification . . . 42
3.1.2 Screening . . . 44
3.1.3 Eligibility . . . 44
3.1.4 Final Inclusion . . . 45
3.2 Descriptive results . . . 45
3.2.1 Corpus overview . . . 46
3.2.2 Publication trends . . . 46
3.2.3 Geography . . . 48
3.2.4 Providers . . . 48
3.2.5 Learning problems and tasks . . . 50
3.2.6 Emerging Role of VLMs . . . 51
3.3 Final considerations . . . 52

4 IDENTIFYING RELEVANT VISUAL ELEMENTS IN URBAN SAFETY PERCEPTION . . . 54
4.1 Introduction . . . 54
4.2 Related Works . . . 55
4.3 Methodology . . . 56
4.3.1 Urban Perception Quantification . . . 57
4.3.2 Urban Elements location using semantic segmentation . . . 57
4.3.3 UrbanFormer Classifier . . . 58
4.3.4 Model Explanation via Grad-CAM and Segmentation IoU . . . 59
4.4 Experiments . . . 61
4.4.1 Dataset Exploration and Preparation . . . 61
4.4.2 Model Training and Performance . . . 62
4.4.3 Model Explanations and Visual Element Importance . . . 64
4.4.4 Limitations . . . 65
4.5 Final Considerations . . . 67

5 COUNTERFACTUAL EXPLANATIONS AND HUMAN-READABLE INTERPRETATIONS OF URBAN SAFETY PERCEPTION . . . 68
5.1 Introduction . . . 68
5.2 Related Works . . . 70
5.3 Methodology . . . 71
5.3.1 Image Segmentation and Manual Annotations . . . 71
5.3.2 Dataset and City Selection . . . 72
5.3.3 Counterfactual Explanations . . . 74
5.3.4 Human-Readable Interpretations via LLM . . . 75
5.4 Experiments . . . 76

5.4.1 UrbanPD4k Dataset and Annotation Statistics . . . 76
5.4.2 Classification Performance and Disorder Impact . . . 77
5.4.3 Counterfactuals and LLM Interpretations . . . 78
5.5 Limitations . . . 79
5.6 Final Considerations . . . 80

6 URBANCOUNTERVIZ: A VISUAL ANALYTICS SYSTEM FOR GROUNDED URBAN SAFETY EXPLORATION . . . 82
6.1 Introduction . . . 82
6.2 Related Works . . . 83
6.2.1 Visual Analytics for Urban Image Analysis and Perception . . . 84
6.2.2 Generative Methods for Urban Scene Editing . . . 84
6.3 System Overview . . . 85
6.3.1 Design Requirements . . . 86
6.3.2 Analytical Tasks . . . 87
6.3.3 Workflow . . . 88
6.4 Data Preprocessing . . . 89
6.4.1 Perception Score Estimation . . . 89
6.4.2 Semantic Feature Extraction and Urban Indicators . . . 89
6.4.3 Discordant Pairs in Urban Perception . . . 90
6.4.4 Similarity Graph Construction . . . 91
6.5 Path-Guided Counterfactual Editing Pipeline . . . 91
6.5.1 Path Guidance . . . 92
6.5.2 Semantic Difference Modeling . . . 92
6.5.3 Two-Stage VLM-Driven Image Editing . . . 93
6.6 The URBANCOUNTERVIZ Interface . . . 97
6.6.1 Source Selector (A) . . . 97
6.6.2 Destination Selector (B) . . . 97
6.6.3 Path View (C) . . . 98
6.6.4 Image View (D) . . . 99
6.6.5 Generation View (E) . . . 99
6.6.6 Image Stats (F) . . . 100
6.7 Implementation Details . . . 100
6.8 Final Considerations . . . 101

7 URBANCOUNTERVIZ EVALUATION . . . 103
7.1 Usage Scenarios . . . 103
7.1.1 Scenario 1 — Enhancing Urban Safety Perception Through Discordant Pairs . . . 104
7.1.2 Scenario 2 — Multi-Step Trajectory Analysis . . . 105

7.1.3 Scenario 3 — Progressive Decrease in Perceived Safety . . . 107
7.2 Expert Feedback . . . 109
7.2.1 Interview Protocol . . . 109
7.2.2 Findings . . . 110
7.2.3 Summary . . . 111
7.3 User Study . . . 112
7.3.1 Study Design . . . 112
7.3.2 Tasks and Procedure . . . 112
7.3.3 Measures . . . 113
7.3.4 Summary . . . 117
7.4 Final Considerations . . . 117

8 DISCUSSION . . . 119
8.1 Synthesis of Findings . . . 119
8.1.1 Grounding Counterfactual Reasoning in Real Urban Data . . . 119
8.1.2 The Role of Semantic Representation . . . 120
8.1.3 Transparency and Human Oversight as System Properties . . . 121
8.1.4 Bidirectionality as a Diagnostic Tool . . . 121
8.2 Limitations . . . 122
8.3 Future Work . . . 125
8.4 Final Considerations . . . 126

9 CONCLUSIONS . . . 128
9.1 Contributions Revisited . . . 128
9.2 Final Remarks . . . 129

REFERENCES . . . 130

1 Introduction

"Cities are designed to shape and influence the lives of their inhabitants" (Lindal and Hartig, 2013). Several studies have shown that the visual appearance of cities plays a central role in human perception and reaction to the environment. A notable example is the "Broken Window theory" (Wilson and Kelling, 1982), which suggests that visual signs of environmental disorder, such as broken windows, abandoned cars, litter, and graffiti, can induce negative social outcomes and increase crime levels. This theory has significantly influenced public policy strategies, leading to aggressive policing tactics to control social and physical disorder manifestations. For example, in the study conducted in the city of New York (Keizer et al., 2008; Hinkle and Weisburd, 2008), social experiments were carried out on the perceived quality of life on the streets. These experiments compared "impeccable" places such as shopping centers (clean walls, orderly and quiet) with other places where there was the presence of graffiti, old or neglected streets, and litter on the streets, concluding that in places where "rules are violated", in the long term, social norms are not respected.

Similarly, other studies have shown that the visual appearance of urban spaces affects the psychological state of its inhabitants (Lindal and Hartig, 2013). For example, a psychological evaluation demonstrated that the presence of green areas in cities tends to produce positive sensations in their inhabitants, such as safety, relaxation, tranquility, etc. (Ulrich, 1979). On the other hand, through the study of 40 psychological reports based on surveys and studies of the mental states of its inhabitants, it was also deduced that urban disorder induces psychological distress, stress, and constant fear (Sampson et al., 2002). In addition, graffiti and buildings in poor or abandoned conditions have also been shown to be directly related to the perception of insecurity, determining that neglected visual appearance negatively influences the environment (e.g., graffiti, litter scattered on the streets, lack of cleanliness in the environment, etc.). (Schroeder and Anderson, 1984).

Therefore, through various studies on the impact of the visual appearance of a city on its inhabitants, it becomes vital to understand the perceptions and evaluations of urban spaces by people. In this sense, several studies have been carried out on how the city and its visual appearance influence the behavior of the city. For example, in "The Image of the City" (Lynch, 1984), cities (such as Boston, Jersey, and Los Angeles) were divided into regions of importance (based on data on crime, society, urban or non-urban sections, etc.), generating mental maps about the common characteristics of these cities. In "The Death and Life of Great American Cities" (Jacobs, 1961), indicates that the elements of each city are distinguished among hundreds, thousands, or millions of other artifacts due to their unique shapes, sizes, colors, etc.

Based on previous studies, a trend began in the psychological aspect of studying and evaluating the perception of the inhabitants regarding the visual elements of the city. In the work carried out by (Nasar, 1998), which was strongly related to finding those regions or areas that were most pleasing to citizens, it was shown that in most evaluations, green areas, themed streets, open spaces, shopping centers, and airports predominated. In addition, areas rated as "not pleasant" for inhabitants were buildings with unattractive styles, graffiti, parks without an established form, and abandoned places. Additionally, other studies related to the disorder of the city (Skogan, 1992) focused on the presence of garbage on the streets, abandoned buildings, and cars parked in desolate corners, which contribute to the perception of lack of control, fear, and insecurity in the city.

Other studies have aimed to understand the correlation between crime behavior and the visual appearance of the city (Skogan, 1992). These studies explore the sense of insecurity linked to the prevalence of crimes and delinquency in public spaces and still rely on in-person evaluations or field observations to collect data on characteristics of the built environment (Rzotkiewicz et al., 2018; Wilson et al., 2012). These studies on crime rates in different cities have led to the development of various applications, including crime maps, data statistics, and applications that predict criminal trends in specific areas (Stalidis et al., 2018), and studies on the impact of visual appearance and crime rate on urban perception in cities (Andersson et al., 2017).

In recent years, with the advancement of tools and frameworks to analyze different types of information (e.g., images, numerical, temporal) and the evolution of techniques such as deep learning, a notable amount of complex databases have been created with the purpose of predicting urban perception. Some works collect data through websites and user online surveys, such as 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), among others. Similar studies quantify greenery and identify green areas and their influence on urban perception (Li et al., 2015b). Furthermore, some studies analyze the relationship between violent levels, the presence of trees and the human development index (Porzi et al., 2015; Arietta et al., 2014; Li et al., 2015a), the presence of graffiti and its correlation with urban safety perception (Tokuda et al., 2019; Lavi. et al., 2022), and divide cities according to the most frequent types of objects (e.g., trees, garbage, buildings, fences, graffiti, etc.) and the respective perception of safety (Zhang et al., 2018b; Min et al., 2020).

Throughout this dissertation, the term *safety* refers to *perceived safety*: a judgment about how safe a street-view scene appears to human observers based on its visual characteristics. It should not be interpreted as an objective measure of crime, physical risk, traffic safety, or actual public safety. This distinction is central to the methodological and ethical framing of the work, since the scores used throughout the dissertation are derived from crowdsourced pairwise comparisons rather than from official crime or incident

records.

However, predicting perceived safety represents only one aspect of the challenge. For effective urban analysis and design, it is also necessary to understand how a scene could plausibly be modified to alter perception. Existing computational approaches typically rely on feature-based explanations or semantic summaries (Pan et al., 2024; Ma et al., 2025; Wu et al., 2025; Min et al., 2020; Song et al., 2025). Although these methods elucidate model behavior, they offer limited support for reasoning about planning-oriented and visually plausible scene modifications. Moreover, counterfactual and generative image-editing approaches, increasingly powered by Vision-Language Models (VLMs), offer a promising direction by illustrating how an image might be transformed to alter a model outcome (Gomez et al., 2020; Augustin et al., 2022; Jeanneret et al., 2022; Hu et al., 2023, 2025; Tan et al., 2025). However, these methods frequently optimize edits based on model responses or open-ended prompt-based generation, rather than grounding them in transformations observed in real urban scenes. Consequently, they may produce visually striking results that are challenging to interpret as realistic urban interventions.

This limitation highlights the need for approaches that connect predictive modeling with interpretable and contextually grounded urban transformations. In particular, analysts require tools that support reasoning about how changes in semantic composition, spatial organization and variation influence perceived safety while remaining consistent with realistic urban conditions. Visualization and visual analytics are well-suited to this challenge because they enable the integrated exploration of image content, semantic structure, model behavior, and domain knowledge. Previous research has demonstrated the value of interactive systems for urban image analysis and machine learning interpretation through comparison, linked views, and human-in-the-loop analysis (Shen et al., 2017; Sacha et al., 2019; Collaris and van Wijk, 2020; La Rosa et al., 2023; Choo and Liu, 2018). However, these approaches primarily support the interpretation of existing data and model outputs, offering limited support for systematically exploring hypothetical yet semantically plausible urban transformations and their predicted effects on perception.

In this work, we introduce UrbanCounterViz (detailed in Chapters 6 and 7), a visual analytics system for the systematic and transparent exploration of urban safety perception through VLMs and semantically guided image counterfactuals. Rather than solely detecting urban elements or relying on open-ended image generation, our approach constructs *grounded counterfactual urban edits* from *semantically meaningful differences* identified between real images with varying perception scores. In this sense, our goal is not only to produce counterfactual examples that change a model output, but to derive recommendations that are supported by real changes and therefore better grounded as analytic evidence.

Before asking how perceived urban safety could be explained or acted upon, it

is first necessary to understand what is already known about it — which is exactly the purpose of the systematic review that opens this dissertation (Moreno-Vera and Poco, 2026). That inquiry reveals a consistent pattern: research has become remarkably good at estimating how safe a street looks, but comparatively little of it explains why, or offers a way to act on that explanation. This gap is where the present work begins.

Understanding perceived urban safety, in this light, is not one question but a sequence of increasingly concrete ones. The first is simply **which visual elements of a street** — its vegetation, its walls, the disorder or upkeep visible in it — **are associated with how safe it is judged to be**. Answering this in a way that can be trusted requires a model that predicts and explains itself at once (Moreno-Vera et al., 2024). But knowing that an element matters says little about how that knowledge could inform plausible changes to the scene: **whether replacing a bare wall with a maintained one, or removing an overhead cable, would plausibly change how a place feels**. Turning that understanding into a concrete, human-readable recommendation is the task studied in Moreno-Vera et al. (2025). And even a well-formed recommendation remains abstract until it can be seen — an analyst reasoning about a street ultimately needs to look at it, not just read about how it might change. Closing that last distance, from a described change to a visible and inspectable one grounded in real urban scenes, is what UrbanCounterViz, introduced in Chapters 6 and 7, sets out to do.

1.1 Thesis statement

Understanding perceived urban safety requires connecting three things that research has largely treated apart: **which visual elements** of a scene shape how safe it feels, **what specific changes** to those elements would plausibly mean, and **whether such changes can be shown** and inspected rather than merely asserted. This dissertation follows that connection from beginning to end, and each chapter that follows takes up the question the previous one left open.

In other words, this thesis argues that perceived urban safety should not be treated only as a prediction problem, but as a grounded counterfactual visual analytics problem: understanding it requires more than accurate predictive models — it requires tools in which models identify perceptually relevant visual elements, translate them into plausible scene changes, and allow analysts to inspect those changes as evidence-based visual transformations.

1.2 Objectives

The objectives of this study are as follows:

Overall objective

This work investigates perceived urban safety from *Street View Imagery* (SVI) through a data-driven pipeline that integrates visual perception modeling, semantic representation, counterfactual reasoning, VLMs, and visual analytics to connect prediction, explanation, and grounded visual intervention.

Specific objectives

The specific objectives of this work follow the same progression as the thesis statement above, moving from mapping what is known, to identifying what matters, to explaining what could change, to making that change visible.

1. **Mapping the state of urban perception research.** Conduct a systematic review of urban perception analysis using SVI, identifying the main approaches, datasets, computational techniques, and methodological gaps in the literature, and use these gaps to motivate the remaining objectives.

2. **Identifying the visual elements that shape perceived safety.** Investigate which specific visual and semantic elements of a SVI are most consistently associated with perceived safety, using a model that predicts and explains its own decisions jointly.

3. **Translating these associations into human-readable recommendations.** Determine what specific changes to a scene's visual elements would plausibly shift its perceived safety, and express these changes as natural-language recommendations rather than as numeric or vector-based output.

4. **Making plausible changes visible and inspectable.** Design and develop a visual analytics system that grounds proposed scene changes in real, observed urban imagery, renders them as photorealistic edits, and allows an analyst to inspect, trace, and interrogate the full transformation process.

1.3 Contributions

This dissertation makes four contributions, each corresponding to one of the objectives above.

Contribution 1. A systematic literature review of computational urban perception research using SVI, identifying four persistent methodological gaps — limited insight into visual factors, static and indirect explanations, unrealistic generative edits, and the absence of interactive human-in-the-loop tools — that motivate the remaining contributions of this thesis. This work was submitted to the journal *IEEE Transactions on Emerging*

Topics in Computational Intelligence (TETCI) as: **Moreno-Vera, Felipe**, and Jorge Poco. "From Vision-centric to Multimodal: A Systematic Review of Street View–Based Urban Perception" (Moreno-Vera and Poco, 2026).

Contribution 2. UrbanFormer (Chapter 4), a classification model that fuses CNN visual features with an explicit semantic composition vector, paired with an explanation methodology that intersects Grad-CAM attention with segmentation masks in image space. This contribution identifies which visual elements most strongly shape perceived safety and establishes the visual vocabulary the remaining contributions build on. This work was published at the *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT '24)* as: **Moreno-Vera, Felipe**, Bruno Brandoli, and Jorge Poco. "What Makes a Place Feel Safe? Analyzing SVIs to Identify Relevant Visual Elements" (Moreno-Vera et al., 2024).

Contribution 3. UrbanPD4k and **UrbanCFX** (Chapter 5): a dataset of 3,659 Rio de Janeiro SVIs with pixel-level annotations of 13 physical disorder elements, and a pipeline that generates counterfactual difference vectors and translates them into human-readable natural language using a *Vision-Language Models* (VLMs). This contribution turns a numeric, model-facing signal into a text-based recommendation a non-specialist can act on directly. This work was published at the *IEEE International Conference on Big Data (BigData '25)* as: **Moreno-Vera, Felipe**, Andres De-la-Puente, and Jorge Poco. "UrbanPhysicalDisorder-4K: Understanding Urban Perception via Counterfactuals and Street View Signs of Physical Disorder" (Moreno-Vera et al., 2025).

Contribution 4. UrbanCounterViz (Chapters 6–7), a visual analytics system that grounds counterfactual scene edits in semantically meaningful differences between real, visually similar urban scenes, and renders them as photorealistic, inspectable transformations through an interactive, human-in-the-loop interface. Alongside the system itself, this contribution includes a reusable interpretable visual analysis workflow that formalizes, independently of any specific implementation, how model predictions, semantic scene changes, and grounded visual transformations can be systematically connected.

During the course of this PhD, we also published additional research that is not directly part of the contributions of this dissertation. In particular, the paper **Moreno-Vera, Felipe**, and Jorge Poco. "Assessing Urban Environments with VLMs: AI-Generated Ratings vs Human Evaluations," published at the *IEEE International Joint Conference on Neural Networks (IJCNN '25)* (Moreno-Vera and Poco, 2025), investigates the use of VLMs for urban perception assessment and complements the broader research agenda pursued during the doctoral program.

Table 1 distinguishes the works that constitute the core contributions of this dissertation from complementary research produced during the doctoral period. The core contributions define the thesis progression, whereas the complementary work informed

Table 1 – Publications produced during the doctorate and their function within this dissertation.

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.

methodological choices but is not developed as an independent chapter.

1.4 Outline

This dissertation is organized into three main parts. The first, comprising Chapters 1 through 3, lays the conceptual and empirical groundwork: it motivates the problem, introduces the technical background needed to follow the rest of the work, and maps the current state of urban perception research through a systematic literature review. The second part, comprising Chapters 4 through 7, presents the three empirical and system contributions developed during the PhD — the systematic review of Chapter 3 constituting the fourth, methodological contribution — following the same progression introduced in this chapter: from identifying which visual elements shape perceived safety, to describing plausible changes to them in plain language, to making those changes visible and inspectable through UrbanCounterViz. The third part, comprising Chapters 8 and 9, draws these contributions together, reflects on their limitations, and outlines directions for future work.

Chapter 2 introduces the technical background — semantic segmentation, counterfactual explanation, and VLMs — needed to follow the contributions presented later. Chapter 3 reports a systematic literature review that maps how urban perception research using SVI has evolved, and identifies the four methodological gaps that motivate this dissertation.

Chapter 4 takes up the first of these gaps, asking which visual elements are most consistently associated with perceived safety. This work was presented at the *IEEE/WIC/ACM International Conference on Web Intelligence* (WI-IAT '24) (Moreno-Vera et al., 2024). Chapter 5 builds on this by translating such associations into concrete, human-readable

recommendations, introducing along the way a dataset of physical disorder annotations for Rio de Janeiro street scenes; this work was presented at the *IEEE International Conference on Big Data* (BigData '25) (Moreno-Vera et al., 2025). A related study, presented at the *IEEE International Joint Conference on Neural Networks* (IJCNN '25) (Moreno-Vera and Poco, 2025), compared AI-generated and human ratings of urban environments using VLMs; while not developed into its own chapter, its findings informed the VLM-based generation strategy adopted in Chapter 6.

Chapters 6 and 7 present UrbanCounterViz, the visual analytics system that closes the distance between a described change and a visible one, grounding every edit in real, observed urban scenes rather than open-ended generation. Chapter 6 situates UrbanCounterViz against the visual analytics and generative editing systems closest to it before detailing its design and implementation, while Chapter 7 evaluates it through usage scenarios, expert interviews, and a broader user study.

Finally, Chapter 8 synthesizes the findings across all contributions and discusses the limitations and future directions of the work as a whole, while Chapter 9 concludes the dissertation by revisiting the objectives set out in this chapter and reflecting on the broader implications of the thesis for urban perception, interpretable AI, and visual analytics.

2 Background

This chapter introduces only the concepts needed to follow the dissertation's contributions: learning tasks for SVI, interpretability and counterfactual explanations, multimodal VLMs, and computational urban perception. Rather than providing a general survey of machine learning, it focuses on the methodological components that are later combined in UrbanFormer (Chapter 4), UrbanCFX (Chapter 5), and UrbanCounterViz (Chapters 6–7). The chapter is organized into four sections. Section 2.1 introduces machine learning paradigms, problem formulations, and learning tasks relevant to SVI analysis. Section 2.2 covers interpretability, gradient-based explanation methods, and counterfactual reasoning — the methodological foundations for the explanation contributions in Chapters 4 and 5. Section 2.3 introduces multimodal representation learning and VLMs, the methodological foundation for the natural-language translation step in Chapter 5 and the two-stage image-editing pipeline in Chapter 6. Section 2.4 introduces urban perception concepts, perception-collection protocols, and quantification methods used throughout this thesis.

2.1 Machine Learning and Deep Learning

Machine learning and deep learning represent two powerful branches of *Artificial Intelligence* (AI) that have revolutionized the way we approach complex problems across various domains. At their core, both machine learning and deep learning aim to enable computers to learn from data and make predictions or decisions without being explicitly programmed for every task (Bengio, 2007).

Machine learning encompasses a range of algorithms and techniques, from classical algorithms like linear regression and decision trees to more advanced approaches such as support vector machines and random forests. Machine learning enables computers to learn patterns and relationships from data and make informed decisions or predictions. Moreover, deep learning has emerged as a particularly powerful approach for tasks that involve complex and high-dimensional data such as images, audio, and text. Deep learning models are built from neural network architectures that learn hierarchical representations directly from data. These neural networks consist of interconnected layers of neurons that process and transform data, learning hierarchical representations of features directly from raw inputs (Goodfellow et al., 2016), through techniques such as CNNs for image recognition, RNNs for sequential data, and transformers for natural language processing. In both cases, we can classify those approaches into four types of learning: supervised, unsupervised, semi-supervised, and reinforcement learning. Besides, each learning type is

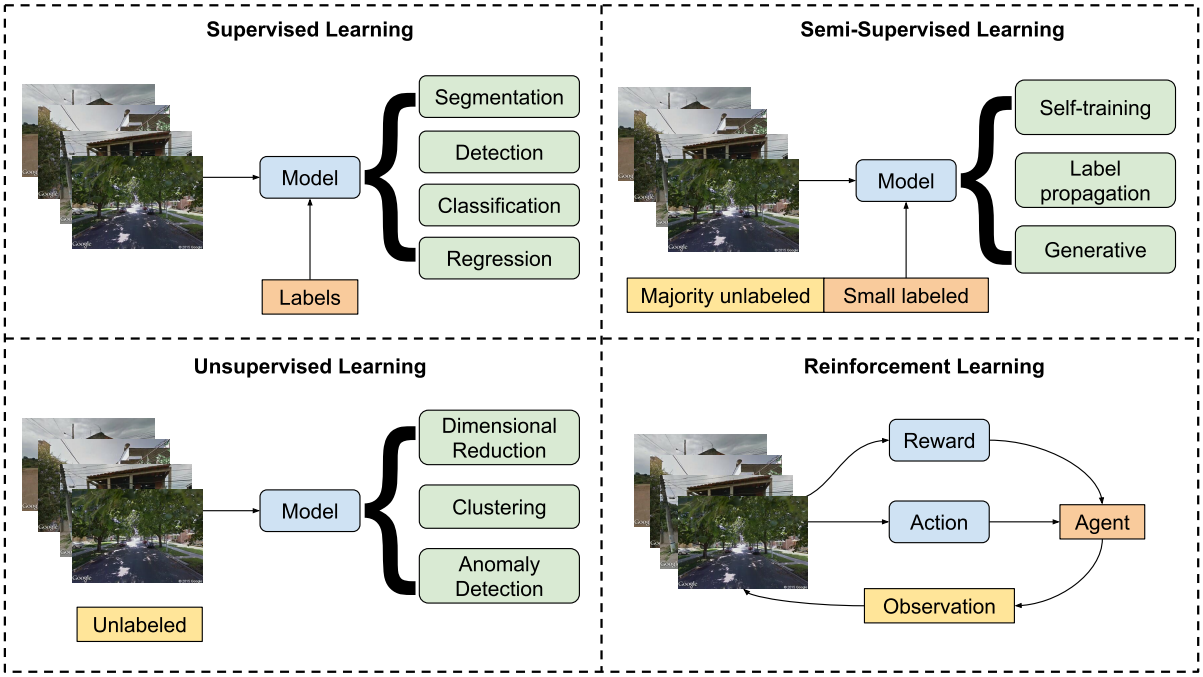

Figure 1 – Main learning problems applied to SVI data as input, we present and specify examples of methods for supervised, unsupervised, semi-supervised, and reinforcement learning settings. Source: the author.

categorized into learning problems and learning tasks.

Learning type

Learning type refers to the overarching approach or paradigm that characterizes how a machine learning algorithm learns from data. The main learning types in machine learning are:

- **Supervised Learning**: In supervised learning, the system is trained on a labeled dataset, where each example in the dataset is paired with a corresponding label or target. The system learns to map input data to output labels by minimizing a predefined loss function, thereby optimizing its predictive accuracy. Examples of supervised learning tasks include classification (assigning labels to inputs) and regression (predicting continuous values).

- **Unsupervised Learning**: Unsupervised learning involves training the system on an unlabeled dataset, where the input data lacks explicit labels or targets. Instead, the system learns to discover patterns, structures, or relationships within the data, such as clusters or latent representations. Common unsupervised learning tasks include clustering (grouping similar data points) and dimensionality reduction (compressing data while preserving its essential features).

- **Semi-supervised Learning**: Semi-supervised learning combines elements of both supervised and unsupervised learning, where the system is trained on a dataset containing a small proportion of labeled examples alongside a larger amount of unlabeled data. The system leverages the labeled data to guide its learning process, while also exploiting the additional information present in the unlabeled data to enhance its performance.
- **Reinforcement Learning**: Reinforcement learning deals with training an agent to interact with an environment in order to maximize cumulative rewards. The agent learns through trial and error, receiving feedback in the form of rewards or penalties based on its actions. By exploring different strategies and learning from the consequences, the agent gradually improves its decision-making capabilities and learns to achieve its goals efficiently.

Figure 1 presents the four types of machine learning and the main learning problems applied to SVI. In this context, supervised learning uses labels to perform classification and regression. Unsupervised learning discovers patterns within unlabeled data. Semi-supervised learning uses a combination of labeled and unlabeled data to improve model performance. Reinforcement learning involves learning through trial and error by interacting with an environment to achieve a goal.

Learning problem

A learning problem refers to a specific objective or task that a machine learning algorithm aims to solve. It defines what the algorithm is expected to learn from the data. Learning problems are typically categorized according to the type of learning involved and the nature of the task. We provide a concise overview of the most commonly used learning problems identified in this work:

- **Classification**: It is a machine learning problem where the goals are to infer a class or category from a set of observations or samples. The input data is $X \in \mathcal{R}^{n \times d}$, where n is the number of samples and d is the number of features (Goodfellow et al., 2016). The output $y \in \{1, 2, \ldots, c\}^n$ is known, where c represents the categories. The aim is to learn a model $f : \mathcal{R}^d \rightarrow \{1, 2, \ldots, c\}$. Classification is a Supervised Learning problem since the input and output are known. Some examples are semantic segmentation, image classification, scene recognition, scene understanding, and object detection.
- **Regression**: Similarly to the classification problem, it is a supervised learning problem where for each sample, we need to estimate a numerical value instead of a class or category (Goodfellow et al., 2016). The model is $f : \mathcal{R}^d \rightarrow \mathcal{R}$, where the input

and output are $X \in \mathcal{R}^{n \times d}$ and $y \in \mathcal{R}$. Some examples are the ridge method, *Least Absolute Shrinkage and Selection Operator* (LASSO), ElasticNet, *Support Vector Regressor* (SVR), and *Support Vector Machine* (SVM), among others.

- **Rank learning**: Also known as learning-to-rank, is a supervised learning task in machine learning that aims to train models to rank a set of items based on their relevance or utility to a given query or user (Sutton and Barto, 2018). Formally, given a dataset $D = (x_i, y_i)_{i=1}^{N}$, where $x_i \in X$ represents the features of the i-th item to be ranked and $y_i \in Y$ represents its ranking. The model is $f : X \to Y$ that maps the input features to the rankings. Some examples are *Ranking Support Vector Machine* (RankSVM), LambdaRank, RankNet, Hinge, and RankBoost.

- **Clustering**: The clustering problem aims to identify groups (called *clusters*) using a set of samples and their features. Unlike the previous problems, clustering is an Unsupervised learning problem since we only have information about the inputs and unknown the outputs (Abu-Mostafa et al., 2012). The aim is to learn a model $f : \mathcal{R}^d \to \{1, 2, \ldots, g\}$ that maps each sample $X \in \mathcal{R}^d$ to a cluster $y \in \{1, 2, \ldots, g\}$. Some samples are Kmeans, *K-Nearest Neighbors* (KNN), and *Gaussian Mixture Models* (GMM), among others.

- **Generative**: In machine learning, Generative refers to a class of models or algorithms that aim to learn the underlying probability distribution of dataset. These models are trained to generate new data samples that are similar to those observed in the training data (Goodfellow et al., 2016). Formally, a generative model G learns the joint probability distribution $P(X, Y)$ of input data X and the corresponding labels or output data Y, where X and Y can be vectors, images, sequences, etc. Given the learned distribution, a generative model can then generate new samples by sampling the learned conditional probability distribution $P(X|Y)$.

- **Policy agent**: It is a machine learning problem in which an agent learns to make decisions by interacting with an environment to achieve a certain goal. In *Reinforcement Learning* (RL), the agent learns through trial and error, receiving feedback in the form of rewards or penalties based on its actions (Sutton and Barto, 2018). Given a set of states S, a set of possible actions, the reward function $R : S \times A \to \mathcal{R}$, and a transition function $T : S \times A \times S \to [0, 1]$ which gives the probability of getting a new state. The aim is to learn an action policy $\pi : S \to A$ which decides which action $a = \pi(s) \in A$ should be taken in order to maximize the final reward after $k \in \mathcal{N}$ steps.

Learning task

A learning task refers to a specific instance or application of a learning problem within a particular context. It often involves defining the input data, the desired output, and the evaluation metrics to assess the performance of the learning algorithm. Learning tasks can vary widely depending on the domain and the specific problem being addressed. Some samples are image classification, object detection, image segmentation, sentiment analysis, anomaly detection, multivariate regression, scene recognition, and ensemble classification, among others.

In summary, learning type describes the overarching approach to learning, learning problem defines the objective or task to be solved, and learning task specifies a particular instance or application of a learning problem in a specific context. Understanding these distinctions helps in effectively framing and addressing machine learning problems in diverse domains.

2.2 Interpretability and Explainability

Interpretability and explainability are related but distinct concepts in the machine learning literature (Lipton, 2018). **Interpretability** is an intrinsic property of a model, referring to the extent to which its internal mechanisms can be directly understood by a human observer. **Explainability**, by contrast, is extrinsic: it concerns the generation of post-hoc artifacts that summarize model behavior in human-understandable terms, without necessarily revealing the underlying computational process. Crucially, a model can be explainable without being interpretable — a deep neural network may produce faithful attribution maps while its learned representations remain opaque.

2.2.1 Explanation Methods

Explanation methods provide insight into the behavior and internal decision-making processes of machine learning models. These methods can be broadly categorized into two families: *white-box* methods, which operate on inherently transparent models such as linear classifiers and are therefore straightforward to analyze, and *black-box* methods, which target complex architectures with large numbers of parameters — such as deep neural networks — where internal reasoning is not directly accessible (Molnar, 2022). Representative techniques in the latter category include **Convolution Visualization** (Zeiler and Fergus, 2014), *smooth* gradients (Ancona et al., 2017), *saliency maps* (Simonyan et al., 2013), and *class activation maps* (CAM) (Zhou et al., 2016), among others.

To formalize the setting, let $x \in \mathbb{R}^d$ denote the feature vector of an input image, $S : \mathbb{R}^d \to \mathbb{R}^C$ a model mapping inputs to scores over C classes, and $E : \mathbb{R}^d \to \mathbb{R}^d$ an explanation function that produces an attribution map over the input dimensions. Under

(a) (b) (c)

Figure 2 – Prediction: Dog; explanations produced by (a) CAM, (b) GBP, and (c) Guided GradCAM. Source: (Selvaraju et al., 2017).

this framework, a gradient-based explanation takes the form $E_{grad}(x) = \frac{\partial S}{\partial x}$, where the gradient captures the sensitivity of each input dimension to the model's prediction $S(x)$ within a local neighborhood of x (Adebayo et al., 2018). The following gradient-based methods build upon this foundation:

- **Gradient Input**: Computes the element-wise product of the input and its gradient, expressed as $x \odot \frac{\partial S}{\partial x}$. This formulation suppresses visual diffusion and mitigates the gradient saturation problem (Shrikumar et al., 2016).
- **Integrated Gradients**: Accumulates gradients along a straight path from a baseline $\bar{x}$ — representing the absence of features — to the input x, yielding $(x - \bar{x}) \int_0^1 \frac{\partial S(\bar{x}+\alpha(x-\bar{x}))}{\partial x} d\alpha$. Integration over the scaling factor α reduces gradient saturation and satisfies a set of desirable axiomatic properties (Sundararajan et al., 2017).
- **CAM (Class Activation Map)**: Produces a class-discriminative localization map of the form $M_{cam} = \sum_k w_k^c F^k$, where w_k^c denotes the weight associated with class c for the k-th convolutional feature map, and $F^k = \frac{1}{Z} \sum_i \sum_j A_{ij}^k$ is the result of applying Global Average Pooling (GAP) to the activation map A^k. Here, (i, j) index spatial positions and Z is the total number of spatial units (Zhou et al., 2016).
- **GradCAM**: Extends CAM by replacing the learned classification weights with gradient-derived importance scores. Specifically, the neuron importance weight is defined as $\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial M_c}{\partial A_{ij}^k}$, referred to as the *partial linearization* of the gradient via GAP. The final localization map is then computed as $M_{gradcam} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$ (Selvaraju et al., 2017).
- **GBP (Guided Backpropagation)**: Modifies standard backpropagation by suppressing negative gradients flowing through ReLU activations. The signal at layer t

for sample i is given by $B_i^t = (f_i^t > 0) \times (B_i^{t+1} > 0) \times B_i^{t+1}$, where $f_i^{t+1} = \text{ReLU}(f_i^t)$ is the forward activation and $B_i^{t+1} = \frac{\partial f^{out}}{\partial f_i^{t+1}}$ is the upstream gradient (Springenberg et al., 2014).

- **Guided GradCAM**: Combines the spatial precision of GBP with the class-discriminative properties of GradCAM through element-wise multiplication of the two attribution maps (Selvaraju et al., 2017).
- **SmoothGrad (SG)**: Reduces noise and visual diffusion in saliency maps by averaging explanations computed over multiple noisy perturbations of the input. For an explanation method E, the smoothed estimate is $E_{sg} = \frac{1}{N}\sum_{i=1}^{N} E(x + g_i)$, where $g_i \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise (Smilkov et al., 2017).

Figure 2 illustrates the outputs of CAM, GradCAM, and GBP on an example image, highlighting the qualitative differences in the attribution maps produced by each method. These techniques are among the most widely used tools for interpreting models trained on visual data.

2.2.2 Interpretability and Counterfactual Methods

While explanation methods focus on understanding *why* a model produces a particular prediction — often through post-hoc attribution of input features — **interpretability** refers to the degree to which a human can understand the internal mechanisms of a model (Lipton, 2018; Doshi-Velez and Kim, 2017). This distinction is subtle but important: a highly accurate deep neural network may be explainable through saliency maps yet remain largely uninterpretable, since the underlying representations encoded in its weights are not directly accessible to human reasoning. Interpretability thus concerns the *transparency* of the model itself, rather than the legibility of individual predictions.

Several lines of research have examined how high-level concepts are represented within neural networks. **Concept-Based Explanations** such as TCAV (Testing with Concept Activation Vectors) (Kim et al., 2018) probe intermediate layers of a network to determine whether human-defined concepts — such as "stripes" or "wheels" — are linearly separable in the learned representation space. This approach allows researchers to quantify the sensitivity of a model's predictions to a given concept without requiring modifications to the model itself. Similarly, **Network Dissection** (Bau et al., 2017) aligns individual hidden units with semantic concepts by measuring the overlap between unit activations and segmentation masks of known visual concepts, revealing that many units in convolutional neural networks correspond to interpretable object parts or textures. A complementary line of work focuses on **Probing Classifiers** (Alain and Bengio, 2018), in which lightweight linear classifiers are trained on the activations of individual layers to

assess what linguistic or semantic properties are encoded at each depth of the network. This approach has been widely adopted in the analysis of VLMs architectures.

Counterfactual explanations offer a different perspective on model behavior by addressing the question: *what is the minimal change to an input x that would alter the model's prediction?* Formally, given a model f and a predicted class $c = f(x)$, a counterfactual x' is sought such that $f(x') \neq c$ and the distance $\|x - x'\|$ is minimized according to some metric (Wachter et al., 2017). Unlike gradient-based attribution methods, counterfactuals do not require access to model gradients and are therefore applicable in black-box settings. In the image domain, **DiCE** (Diverse Counterfactual Explanations) (Mothilal et al., 2020) generates multiple diverse counterfactuals to prevent over-reliance on a single explanation, while **Contrastive Explanations** (Dhurandhar et al., 2018) decompose explanations into *pertinent positives* (features that must be present to justify the prediction) and *pertinent negatives* (features whose absence is necessary).

A particularly relevant formulation for visual models is that of **semantic counterfactuals**, in which changes are made in a latent semantic space rather than in pixel space, yielding more coherent and human-interpretable modifications (Goyal et al., 2019). This line of work has been applied to visual perception tasks where understanding the role of object attributes — such as color, texture, or spatial configuration — is important for auditing model decisions.

Building on this idea, this dissertation adopts the notion of a **grounded counterfactual**: a proposed change to an urban scene whose direction is derived from differences observed between real SVIs, rather than from unconstrained prompt-based generation or purely synthetic optimization in feature or latent space. Grounding therefore refers to the empirical origin of the edit — the transformation is anchored in visual and semantic patterns that actually occur in the data. This formulation, central to the system presented in Chapter 6, trades some of the flexibility of open-ended generation for plausibility and traceability: every proposed change can be traced back to an observed difference between real scenes.

Together, interpretability methods and counterfactual explanations complement the gradient-based attribution approaches described in the previous section, providing a more complete picture of model behavior: whereas attribution methods highlight *where* a model looks, interpretability methods reveal *what* it has learned, and counterfactual methods characterize *how* it would decide differently — with grounded counterfactuals additionally constraining that *how* to changes observed in real data.

2.3 Multimodal Learning and Vision-Language Models (VLMs)

The methods discussed so far operate on a single input modality, typically an image, and produce either a class label, a continuous score, or an attribution map defined over that same modality. Many of the contributions in this thesis, however, require reasoning that spans two modalities simultaneously: an urban scene expressed as an image and a description of that scene, or of a change to that scene, expressed in natural language. This section introduces the concepts of multimodal representation learning and VLMs (VLMs) that underlie the natural-language translation of counterfactual changes in Chapter 5 and the vision-language image-editing pipeline of Chapter 6.

2.3.1 Multimodal Representation Learning

Multimodal learning refers to a family of methods that jointly process information from more than one modality — such as vision, language, and audio — in order to learn representations that capture relationships across them (Baltrušaitis et al., 2018). Formally, let $x_I \in \mathcal{X}_I$ denote an image and $x_T \in \mathcal{X}_T$ denote a text sequence. A multimodal model learns a pair of encoders,

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

that map inputs from each modality into a shared embedding space of dimensionality d, such that semantically related image-text pairs are placed close together under a similarity function $\mathrm{sim}(\cdot, \cdot)$, typically cosine similarity,

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

This shared embedding space is what allows a downstream model to compare, retrieve, or condition generation across modalities, and constitutes the basic building block of the VLMs described next.

2.3.2 Vision-Language Models (VLMs)

VLMs extend the multimodal formulation above by training the image encoder f_I and text encoder f_T jointly on large corpora of paired image-text data, most commonly using a contrastive objective (Radford et al., 2021). Given a batch of N image-text pairs $\{(x_I^i, x_T^i)\}_{i=1}^N$, the model is trained so that the embeddings of matched pairs are pulled together while embeddings of mismatched pairs are pushed apart. This is achieved with the InfoNCE loss (Oord et al., 2018), applied symmetrically over images and text,

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

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

where τ is a learned temperature parameter controlling the sharpness of the distribution. The overall training objective is the average $\mathcal{L}_{\text{VLM}} = \frac{1}{2}\left(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}\right)$. Models trained under this contrastive formulation, such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), learn representations that transfer well to downstream tasks — including zero-shot classification and cross-modal retrieval — without task-specific fine-tuning, since a novel class or query can be expressed simply as a text prompt and compared against image embeddings via Equation (2.2).

A complementary line of work replaces the pure contrastive objective with an encoder-decoder or unified transformer architecture trained on a mixture of contrastive, captioning, and matching objectives, such as BLIP (Li et al., 2022a) and BLIP-2 (Li et al., 2023). These models pair a frozen or lightly-tuned vision encoder with a language decoder, enabling not only alignment between modalities but also the generation of free-form text conditioned on visual input — a capability that contrastive-only models such as CLIP lack, and that the generative VLMs described next build upon.

2.3.3 Generative and Instruction-Following VLMs

While the models above focus on *aligning* the two modalities, a second family of VLMs focuses on *generation*: producing free-form text conditioned on visual input, following a natural-language instruction. Given an image x_I and an instruction or prompt x_T, these models are typically built on an autoregressive language model that generates an output sequence $y = (y_1, \ldots, y_L)$ token by token, conditioned on both inputs,

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

where the image representation $f_I(x_I)$ — typically a sequence of visual tokens or patch embeddings projected into the language model's embedding space — is treated as additional context alongside the previously generated tokens $y_{<t}$. Training proceeds by minimizing the negative log-likelihood of a reference response, following the same cross-entropy formulation introduced for the classification setting in Section 2.1, but summed over the sequence:

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

Representative models trained under this paradigm include Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023a), and large-scale proprietary systems such as GPT-4V (Achiam et al., 2023) and Gemini (Team et al., 2023), which additionally support multi-turn, instruction-following interaction. In the context of this thesis, this generative formulation is the mechanism by which a numeric or symbolic change to an urban scene — for instance,

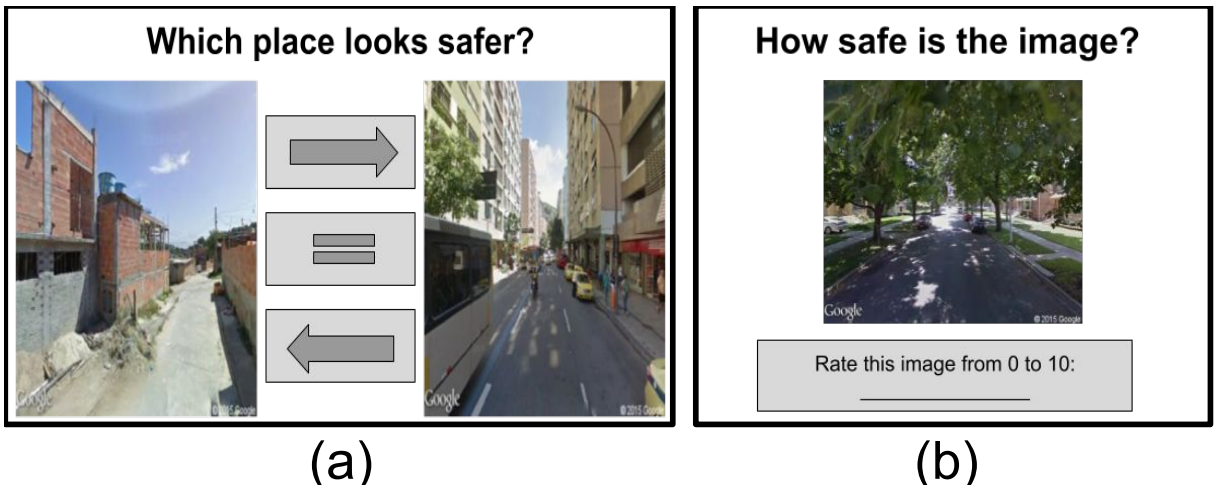

Figure 3 – There are two main collection approaches: (a) gather volunteer information through surveys, interviews, or questionnaires asking e.g., *Which place looks safer?*, which volunteers choose between two images; or (b) "Rate the safety of this image" ask for volunteers to rate an image from 0 to 10. Source: the author.

a counterfactual difference vector — is translated into a natural-language recommendation in Chapter 5, and by which an editing instruction is subsequently converted into a photorealistic image edit in the two-stage pipeline of Chapter 6, where a VLMs conditions on both the source image and a structured description of the intended change to produce the edited scene.

2.4 Urban Perception and AI

Street View Imagery provides large-scale visual coverage of urban environments, and machine learning methods — particularly deep neural networks for object detection, semantic segmentation, and scene understanding — make it possible to analyze this imagery at scale (Ibrahim et al., 2020; Zhang et al., 2024). This combination underlies the field of computational urban perception, whose data collection and quantification protocols this section introduces.

In general, urban perception studies aim to collect and quantify human perceptions of urban environments. Many of these studies rely on web-based platforms, such as Place Pulse (Salesses et al., 2013), City-SAFE (Costa et al., 2019), and Wmodi (Acosta and Camargo, 2018), to gather large-scale perception data. Participants are typically presented with pairs of SVIs and asked to compare them according to specific attributes, such as safety, beauty, liveliness, or wealth. The resulting datasets generally include the image pairs, their geographic coordinates, and the corresponding human judgments. These annotations can then be used to train machine learning and deep learning models capable of predicting perceptual attributes directly from urban imagery, enabling large-scale assessments of urban environments without requiring continuous human evaluation.

Four terms recur throughout this dissertation and deserve explicit definition. **Perceived safety** denotes a judgment about how safe a street-view scene appears to human observers, derived from crowdsourced pairwise comparisons; it is distinct from objective safety, crime rates, or physical risk (see Chapter 1). A **semantic representation** of a scene is a structured description of its content in terms of labeled categories — for example, the per-category pixel-ratio vectors produced by semantic segmentation — rather than raw pixel values. **Visual analytics** refers to analytical reasoning supported by interactive visual interfaces, in which computational models and human judgment are combined through coordinated views rather than static outputs. Finally, a system is **human-in-the-loop** when the analyst actively participates in the analytical process — selecting inputs, steering transformations, and validating outputs — instead of receiving a single, non-negotiable result. The related notion of a *grounded counterfactual* is defined in Section 2.2.

2.4.1 Perception collection

Figure 3 illustrates two main methods for collecting perceptual judgments: (a) rating scales and (b) ranking pairwise comparisons.

Rating scores, commonly using the Likert scale Likert (1932), ask participants to rate images on a fixed, ordinal scale (e.g., 5 or 10 points) with options ranging from *strongly agree* to *strongly disagree*. This approach captures varying degrees of perception or preference, allowing for more nuanced responses than simple binary choices. As a result, researchers can better quantify subjective impressions and compare results across participants. The final image score is typically computed as the average of all responses, known as the **Mean Score**, which provides a straightforward summary of overall evaluation.

Ranking scores, particularly pairwise comparisons Thurstone (2017), present two images and ask users to select the preferred one based on a given criterion (e.g., safety, beauty, or attractiveness). This approach enables a more granular understanding of preferences, as individuals make direct comparisons rather than evaluating images in isolation. From these comparisons, a final **ranking score** for each image can be estimated using methods such as the "strength of schedule" algorithm Park and Newman (2005); Naik et al. (2014); Salesses et al. (2013), which accounts for win–loss rates and opponent difficulty; the Elo rating system Ebtekar and Liu (2021), which updates scores iteratively based on pairwise outcomes; and TrueSkill Minka et al. (2018), a Bayesian framework that models both skill estimates and uncertainty.

2.4.2 Perception calculation

While ratings provide scores directly, pairwise comparison is a common strategy for quantifying subjective perceptions from visual data. In this approach, annotators

select which image better represents a given attribute (e.g., safer, more beautiful, or more lively). These binary outcomes can then be used by ranking algorithms to estimate latent perceptual scores for each image.

Let i and j denote two images compared in a pairwise task. The comparison outcome is represented as:

$$S_i = \begin{cases} 1, & \text{if image } i \text{ is preferred over image } j \\ 0, & \text{otherwise} \end{cases} \tag{2.7}$$

Similarly, the outcome for image j is defined as:

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

Elo Rating System

The Elo rating system models the probability that one image is preferred over another as a function of the difference in their ratings. Each image is assigned a scalar rating R, which is iteratively updated after each comparison. The expected probability that image i is preferred over image j and the expected probability for image j are defined as:

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

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

After observing the comparison outcome, the ratings are updated according to:

$$R_i' = R_i + K\,(S_i - E_i) \tag{2.10a}$$

$$R_j' = R_j + K\,(S_j - E_j) \tag{2.10b}$$

where K is a constant controlling the learning rate of the updates. Typical values of K range from 16 to 32. The Elo method produces a single deterministic score for each image and is computationally efficient for large-scale pairwise datasets.

TrueSkill Rating System

The TrueSkill framework (Minka et al., 2018) extends the Elo system by modeling uncertainty in the estimated score of each image. Instead of a single rating value, each image is represented by a Gaussian distribution:

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

where μ_i represents the estimated skill (or perception score) and σ_i represents the uncertainty associated with that estimate. TrueSkill assumes that the observed performance of an image in a comparison is subject to noise:

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

where β is a parameter representing performance variability. For a pairwise comparison in which image i is preferred over image j, the combined variance is computed as:

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

The standardized performance difference is then defined as:

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

Two auxiliary functions are used in the update equations:

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

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

where $\phi(t)$ and $\Phi(t)$ denote the probability density function (PDF) and cumulative distribution function (CDF) of the standard normal distribution. The updated mean values are computed as:

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

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

The updated variances are:

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

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

In practice, images are ranked using a conservative estimate of their score:

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

This formulation penalizes high uncertainty and produces stable rankings, particularly in crowdsourced perception datasets with varying numbers of comparisons per image.

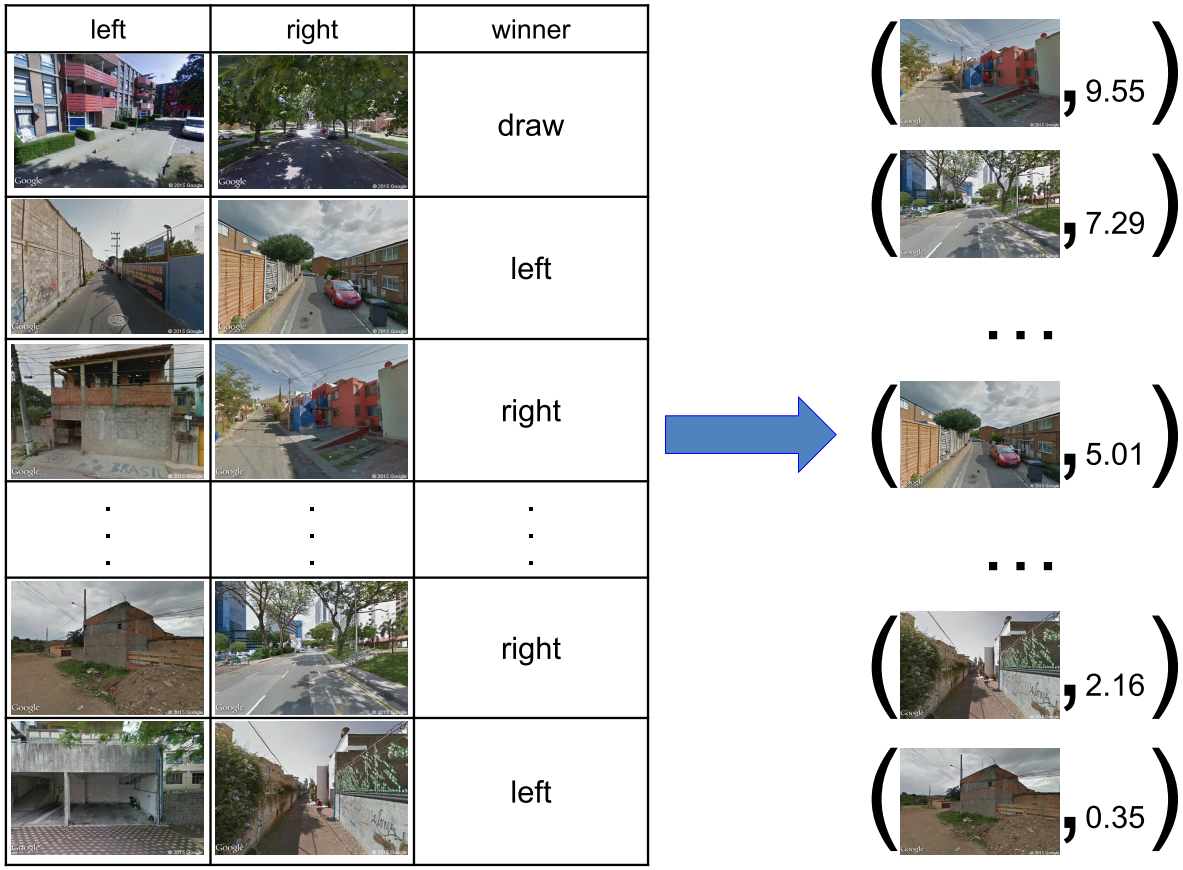

Figure 4 – Quantifying the human evaluations of urban perception using SVI, from the comparison between two images it is possible to calculate scores for each image evaluated. Source: the author.

Strength of Schedule

Strength of Schedule (SOS) (Park and Newman, 2005) is a method used to evaluate the difficulty of a team's opponents, helping to provide a fairer assessment of performance. It typically considers the winning percentages of opponents and, in some cases, the opponents of those opponents. This metric is widely used in sports rankings to compare teams with different levels of competition. So, we can calculate rates, awards, and penalty for each image:

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

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

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

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

Where w_i^k, d_i^k, and l_i^k denote the number of times image i was chosen as the winner,

tied, or loser; n_1 and n_2 are the win and loss counts; *Award* is the average win rate when image i won, and *Penalty* is the average loss rate when it lost. The **Q-scores** are then scaled from 0 (very unsafe) to 10 (very safe). Figure 4 illustrates how pairwise comparisons can be transformed into quantitative measures, enabling the assignment of scores or rankings to images.

The SOS formulation presented here is the canonical reference for all perception score computations in this thesis; subsequent chapters refer back to this section rather than re-deriving it independently.

2.4.3 Perception tasks

Urban perception modeling is predominantly framed as a supervised learning problem, where models are trained to predict human judgments collected through surveys based on SVI. Depending on the nature of the target variable and the available annotations, existing studies typically formulate the problem according to one of three learning paradigms.

(i) Classification treats urban perception as a categorical prediction task, assigning discrete labels (e.g., *safe* or *unsafe*) to urban scenes. Given an input urban scene $\mathbf{x}$, a classifier predicts the probability distribution over K perception classes,

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

where $f_\theta(\cdot)$ denotes the model parameterized by θ. The model is commonly optimized using the cross-entropy loss,

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

where y_k is the one-hot encoded ground-truth label. Classification models often leverage visual features extracted from SVI and may incorporate complementary sources of information, including demographic indicators, geospatial attributes, and textual descriptions, to improve predictive performance.

(ii) Regression aims to estimate continuous perceptual scores derived from aggregated human ratings. Instead of predicting discrete categories, the model outputs a scalar score,

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

where $\hat{s} \in \mathbb{R}$ represents the predicted perception score. Training is typically performed by minimizing the mean squared error (MSE),

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

where s_i denotes the ground-truth perception score. This formulation enables a more nuanced representation of subjective perceptions by capturing varying degrees of attributes. Regression-based approaches are particularly useful when perception scores are measured on numerical scales and fine-grained predictions are desired.

(iii) Ranking and pairwise learning focuses on learning relative preferences between urban scenes rather than absolute labels or scores. Given a pair of urban scenes $(\mathbf{x}_i, \mathbf{x}_j)$ with corresponding latent scores r_i and r_j, the probability that scene i is preferred over scene j can be modeled as

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

where $\sigma(\cdot)$ is the sigmoid function. The corresponding pairwise ranking loss is:

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

Models are typically trained using ranking losses or probabilistic frameworks, including Elo- and TrueSkill-based formulations, to infer latent perceptual rankings from pairwise comparisons.

Although these formulations differ in their prediction objectives and supervision signals, they all rely on learning representations that capture the visual and contextual characteristics underlying human perception of urban environments.

2.4.4 Perception dataset

The experiments throughout this thesis draw on a single large-scale perception dataset. We use the Place Pulse 2.0 dataset, which in its original release consists of approximately 1.22 million raw pairwise image comparisons involving 111,390 SVIs from 56 cities across 28 countries, collected using *Google Street View* (GSV). The images were annotated across six perceptual categories: *safety*, *liveliness*, *beauty*, *wealth*, *depression*, and *boredom*. Each comparison record provides the image identifiers, geographic coordinates, and the identity of the winning image for the corresponding perceptual category. **The released dataset does not include per-annotator information such as nationality or demographics; consequently, no personal characteristics of the annotators (volunteers) can be analyzed.**

To transform pairwise comparison outcomes into quantitative perceptual scores, we apply the SOS algorithm (Park and Newman, 2005) introduced in Section 2.4.2. Instantiating the canonical formulation (Equation (2.19d)) for the *safety* category (k), the perceptual Q-score of each image i is:

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

where Rate_i^k, Award_i^k, and Penalty_i^k are defined as in Section 2.4.2, and the resulting score is scaled to the interval $[0, 10]$, where a value near zero indicates very unsafe perception and a value near ten indicates very safe perception. Table 2 reports the statistics of the subset retained after pre-processing the comparisons.

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 – Statistics obtained after process all comparisons from Place Pulse, containing information about images per city in each continent (a) and the mean score for each requested category (b).

2.5 Final Considerations

This chapter introduced and described the main concepts in machine learning and deep learning commonly applied to urban perception analyses using SVI. Beginning with classical machine learning algorithms and linear models, and progressing to deep learning architectures, we covered the principal learning paradigms, problem formulations, and tasks relevant to the processing of SVI. We further distinguished between interpretability and explainability, surveying the most widely used attribution and counterfactual methods for understanding model behavior. We then introduced multimodal representation learning and VLMs, from contrastive image-text alignment to generative, instruction-following architectures, which underpin the language-based and image-editing components of the contributions presented later in this thesis. Finally, urban perception concepts and evaluation frameworks were introduced to contextualize the contributions presented in the following chapters. With these foundations in place, Chapter 3 presents a systematic literature review of how these methods have been applied in the urban perception field. The review covers 234 studies spanning classification, regression, ranking, segmentation-based, and multimodal approaches to urban perception, and concludes with four methodological gaps that directly motivate the empirical contributions and system design presented in the remainder of this thesis.

3 Systematic Literature Review

As motivated in Chapter 1, a central challenge in urban perception research is that existing approaches predominantly focus on prediction accuracy while offering limited insight into the visual and semantic factors that drive human judgments of urban safety. To systematically characterize the state of the field and identify the methodological gaps that this thesis addresses, we conducted a structured review of the literature on urban perception using SVI.

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) protocol (Regona et al., 2022)[1]. PRISMA is a widely recognized framework that enhances the transparency, consistency, and thoroughness of systematic reviews. It has been used in related studies (Charreire et al., 2014; Yin et al., 2023; Marasinghe et al., 2023; He and Li, 2021). The method is divided into four steps: identification, screening, eligibility, and inclusion. The review concludes by identifying four persistent methodological gaps in the literature — limited insight into visual factors, static and indirect explanations, unrealistic generative edits, and the absence of interactive human-in-the-loop tools — each of which is directly addressed by one or more contributions of this thesis.

3.1 Methodology

Figure 5 illustrates this review process. We first selected relevant keywords to identify an initial set of papers, which were screened to filter out irrelevant works. We then reviewed the content of the remaining papers, focusing on studies related to urban perception using SVI from 2005 to 2026. The starting year was selected because of the launch of major street-view platforms, including Baidu Street View (2005), Google Street View (2007), and Tencent Street View (2011). During the eligibility and inclusion stages, we employed snowballing (Wohlin, 2014) to discover additional relevant research not identified in the initial search. Further details of this process are described in the following sections.

3.1.1 Identification

This step, also known as the *search criteria*, aimed to identify an initial set of papers. We searched ScienceDirect, Google/Semantic Scholar, and Web of Science using terms such as *street view*, *street view imagery*, *street imagery*, and *street level*. Additional keywords, including *Measuring urban*, *Quantifying urban*, *Safety perception*,

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

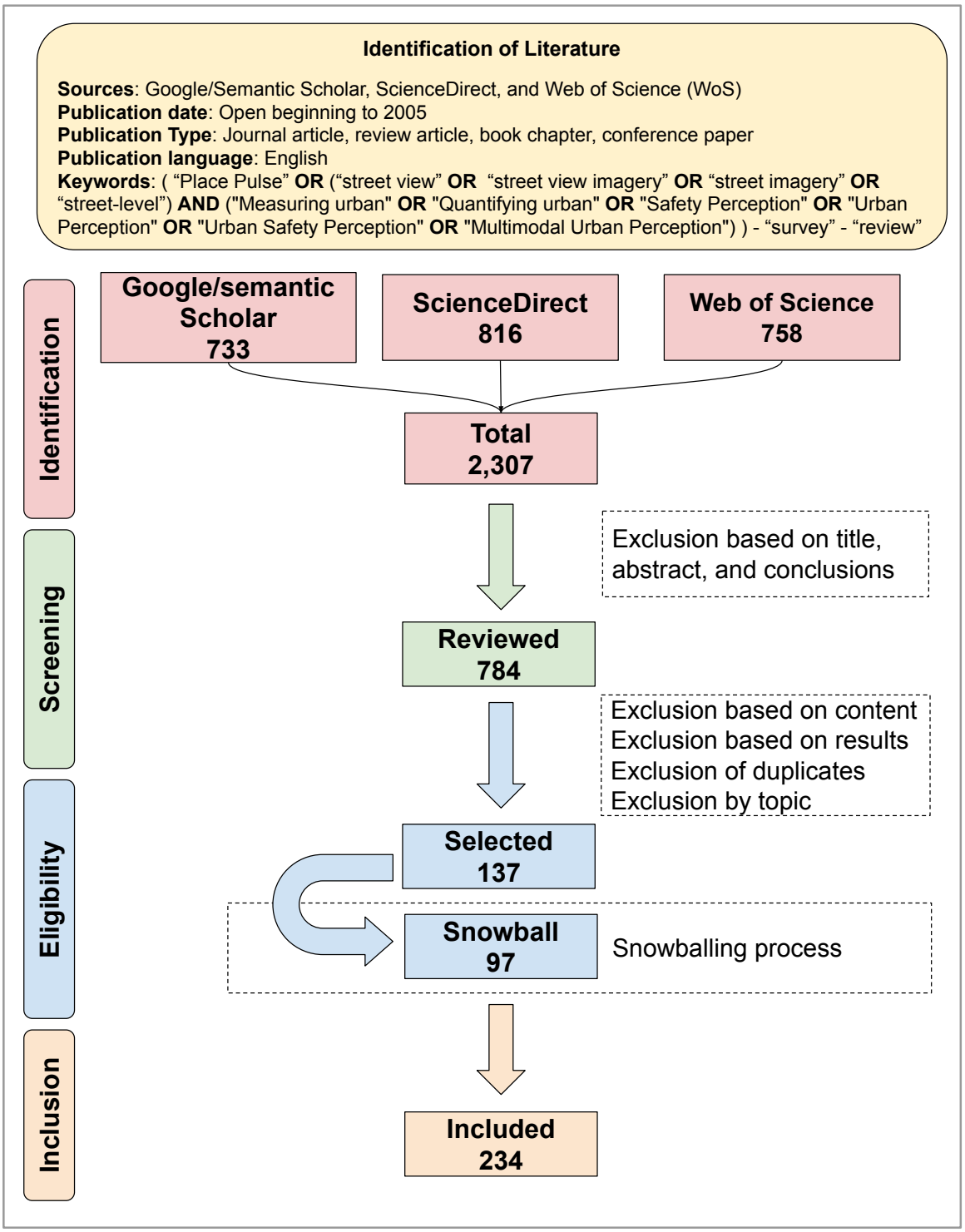

Figure 5 – Flow of the systematic review according to the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) protocol. Note that we also include snowballing in the last step of the subprocess to review and capture more articles for this review. Source: the author.

Urban Perception, *Multimodal Urban Perception*, and *Urban Safety Perception*, were used to capture studies on urban perception using SVI. The term *Place-Pulse* was also included to identify studies that followed that dataset's methodology. Table 3 shows the number of

Table 3 – Number of papers per step and keywords

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

papers identified and retained throughout the review process after duplicate removal.

Although terms like *street view* and *street* are generic, they help retrieve papers referencing services such as *Google Street View*, *Tencent Street View*, *Naver Street View*, *Baidu Street View*, and *OpenStreetMap*, among other services. The search was expanded to include related mapping services: combining *mapping*, *street view*, and *urban perception* yielded 39 results, while excluding *urban perception* produced 1,470. Irrelevant studies (e.g., pollution, thermal inference, remote sensing, aerial environments, temporal change, and weather analysis) were excluded. While *Street View* returned more results, most were excluded; *Urban perception* and *Place-Pulse* yielded the most relevant studies. In total, this step identified **2,307 papers**.

3.1.2 Screening

This step, also called *selection criteria*, involved reviewing the titles, abstracts, data, methods, and conclusions of the initial 2,307 papers to identify a relevant subset. The inclusion criteria were: (1) English language; (2) focus on urban perception in a street-level context; (3) goes beyond computer vision by incorporating approaches such as statistical analysis and deep learning; and (4) examines human perception preferences collected through SVI and online surveys. **We excluded** topics such as weather and climate, pollution, population studies, solar and thermal analysis, agricultural monitoring, traffic, and city simulation, as well as studies focused solely on computer vision or deep learning tasks (e.g., image segmentation models in urban scenes). This step resulted in a total of **784 papers**.

3.1.3 Eligibility

We reviewed 784 papers by extracting key characteristics, including SVI service, publication year, keywords, venue, author affiliations, datasets, study location, machine learning task, references, and citations. Papers were excluded if they were (1) duplicates, (2) non-peer-reviewed (e.g., abstracts, presentations, books, case reports, arXiv or similar

preprints), or (3) unrelated to urban perception. This process resulted in **137 papers**. We then applied snowballing to identify additional relevant studies.

Snowballing process

Snowballing, or citation chaining, expands relevant literature beyond initial search results Wohlin (2014). It involves iteratively examining references of selected papers and identifying new relevant studies through *backward* and *forward* steps:

- **Backward snowballing** starts with the references of identified papers and works backward.
- **Forward snowballing** explores the papers that have cited the identified papers, moving forward in time.

We used Semantic Scholar[2] to identify highly influential references and citations, based on contextual signals and citation frequency. In both steps, we also verified language, geographic origin, and venue type, following the same filtering criteria as in the prior screening step. Applying this to the initial 137 papers, we selected 50 additional papers in the backward step and 47 in the forward step, resulting in a total of **234 papers included in this review**.

3.1.4 Final Inclusion

From an initial set of 2,307 papers, 234 met the inclusion criteria. These studies were systematically analyzed and categorized using a four-component multimodal framework that reflects the typical data-driven pipeline in urban perception research, from data collection and integration to analysis and real-world application, while incorporating a multimodal approach. Given their breadth and diversity, we believe this review captures key trends, methods, and emerging directions while minimizing selection bias. The full list of included papers is available in a public repository.[3]

3.2 Descriptive results

Here, we present the principal descriptive findings obtained from the *Systematic Literature Review* (SLR).

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

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

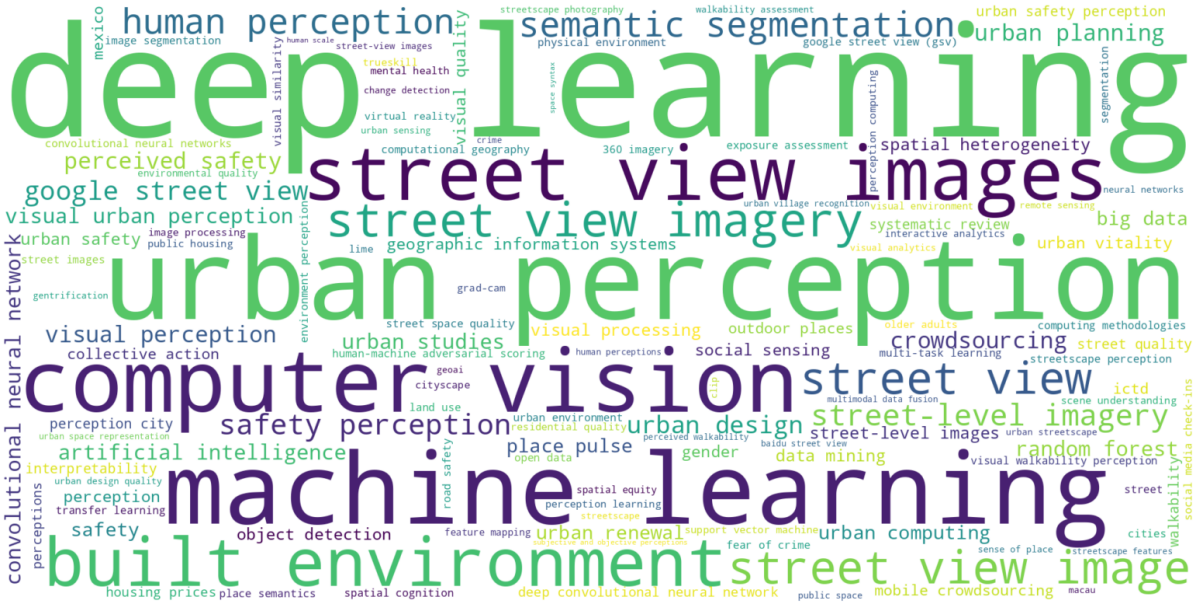

Figure 6 – Word cloud depicting the most frequently occurring keywords from 234 studies included. Source: the author.

3.2.1 Corpus overview

The 234 selected studies were published across 113 distinct venues, underscoring the inherently interdisciplinary character of urban perception research. The journals with the highest representation are *Cities* (16 papers), *Landscape and Urban Planning* (12 papers), *Computers, Environment, and Urban Systems* (11 papers), and *ISPRS International Journal of Geo-Information* (10 papers). This distribution illustrates the intersection of several research domains, including urban studies, geographic information science, computer vision, remote sensing, and artificial intelligence.

Keyword analysis further highlights the methodological orientation of the field. The most frequently reported terms include *deep learning*, *urban perception*, *machine learning*, *computer vision*, and *street view images*. Together, these keywords indicate a strong reliance on image-based representations and AI-driven techniques for analyzing and modeling human perceptions of urban environments. The complete keyword frequency distribution is presented in Figure 6.

3.2.2 Publication trends

Although the search period began in 2005, all studies included in this review were published after 2012. This period coincides with both the rise of convolutional neural networks (CNNs) and the first large-scale efforts to quantify urban perception through online crowdsourcing platforms Salesses et al. (2013), marking the emergence of street-view imagery–based urban perception research as a distinct field. Early contributions (2012–2014) predominantly employed handcrafted visual descriptors, including Gist features, color

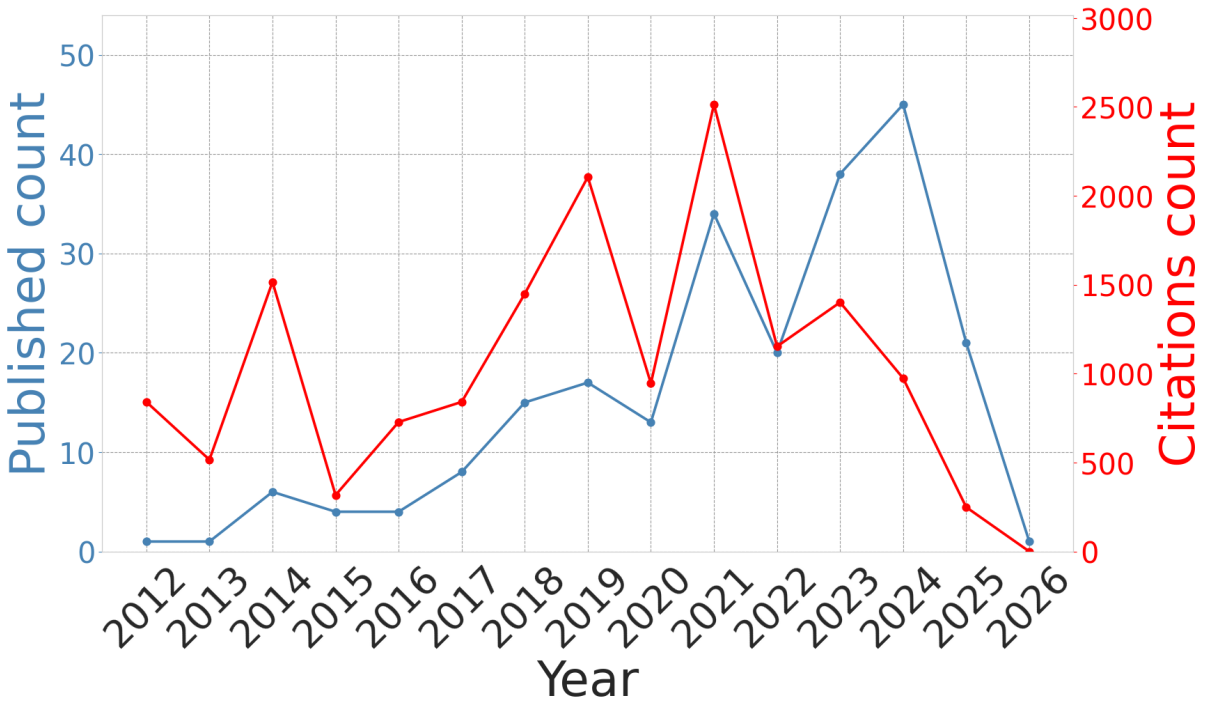

Figure 7 – The number of published papers (blue) and citations (red) by year shows a clear growth in research output on urban perception using SVI, particularly after 2016, with publication counts peaking between 2021 and 2024. Citations exhibit a more fluctuating pattern, reaching their highest levels around 2019 and 2021, followed by a noticeable decline in recent years. The sharp decrease in citations for 2024–2026 is expected, as these publications are more recent and have had less time to accumulate citations. Source: the author.

histograms, and texture-based representations Doersch et al. (2012); Wilson et al. (2012); Lindal and Hartig (2013); Ordonez and Berg (2014); Naik et al. (2014). As deep learning matured, these approaches were progressively replaced by CNN-based architectures and, in some cases, reinforcement learning techniques Wang et al. (2022a,b).

More recently, the field has experienced a new methodological shift driven by the emergence of *Large Language Models* (LLMs) Zhang et al. (2025a); Malekzadeh et al. (2024); Huang et al. (2024); Levering et al. (2024) and multimodal models (mainly VLMs) Moreno-Vera and Poco (2025); Zhang et al. (2025b); Han and Kim (2025), which enable richer multimodal reasoning and more flexible representations of urban scenes.

Figure 7 reports annual publication counts together with cumulative citation totals by publication year, obtained through the Semantic Scholar API. Publication activity remained relatively stable during 2014–2016 and 2018–2020 but accelerated considerably after 2020. This growth appears to be associated with the increasing availability of large-scale SVI datasets, the incorporation of complementary urban data sources, and the widespread adoption of foundation models and multimodal learning techniques.

It is important to note that citation values correspond to the total number of citations accumulated by papers published in a given year rather than citations received

during that year. Consequently, recent publications naturally exhibit lower citation counts due to their shorter exposure time. As of April 2026, the studies included in this review had collectively accumulated 16,967 citations. Five publications from 2026 were also included, explaining the absence of citations for that year.

3.2.3 Geography

Analysis of author affiliations reveals a strong concentration of research activity in a small number of countries. China accounts for 35.43% of all affiliations, followed by the United States (17.21%) and the United Kingdom (8.60%). However, the thematic emphasis of urban perception research varies across regions.

Research conducted by U.S.-based institutions primarily focuses on improving predictive models and developing novel computational methods for perception estimation. In contrast, Chinese institutions often emphasize both methodological development and large-scale applications to rapidly urbanizing cities, frequently adapting or extending benchmark datasets such as Place Pulse 2.0 Dubey et al. (2016). UK-based studies commonly investigate the relationships between urban form, aesthetics, and human perception, including assessments of visual attractiveness and environmental quality 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).

The growth of Chinese research output between 2016 and 2019, visible in Figure 7, coincides with the broader adoption of deep learning techniques and the increasing accessibility of large-scale Chinese street-view datasets.

Although research production is dominated by China, the United States, and the United Kingdom, the reviewed literature spans 41 additional countries. Regarding study locations, China (10.35%), the United States (5.89%), and the United Kingdom (4.28%) are also the most frequently investigated settings. Affiliation and study-country percentages were computed at the paper level using all authors' institutional affiliations and the geographical focus of each study. Figure 8 illustrates these relationships through a Sankey diagram, highlighting China as the most prominent study location across the corpus.

3.2.4 Providers

Street View Imagery (SVI) has become a fundamental data source for urban perception research because it enables large-scale virtual observation of streetscapes and urban environments Li et al. (2022b). In many studies, road networks are first obtained from sources such as OpenStreetMap (OSM), GeoGraph, or GIS platforms before imagery is systematically collected from street-view services. Among the providers identified in

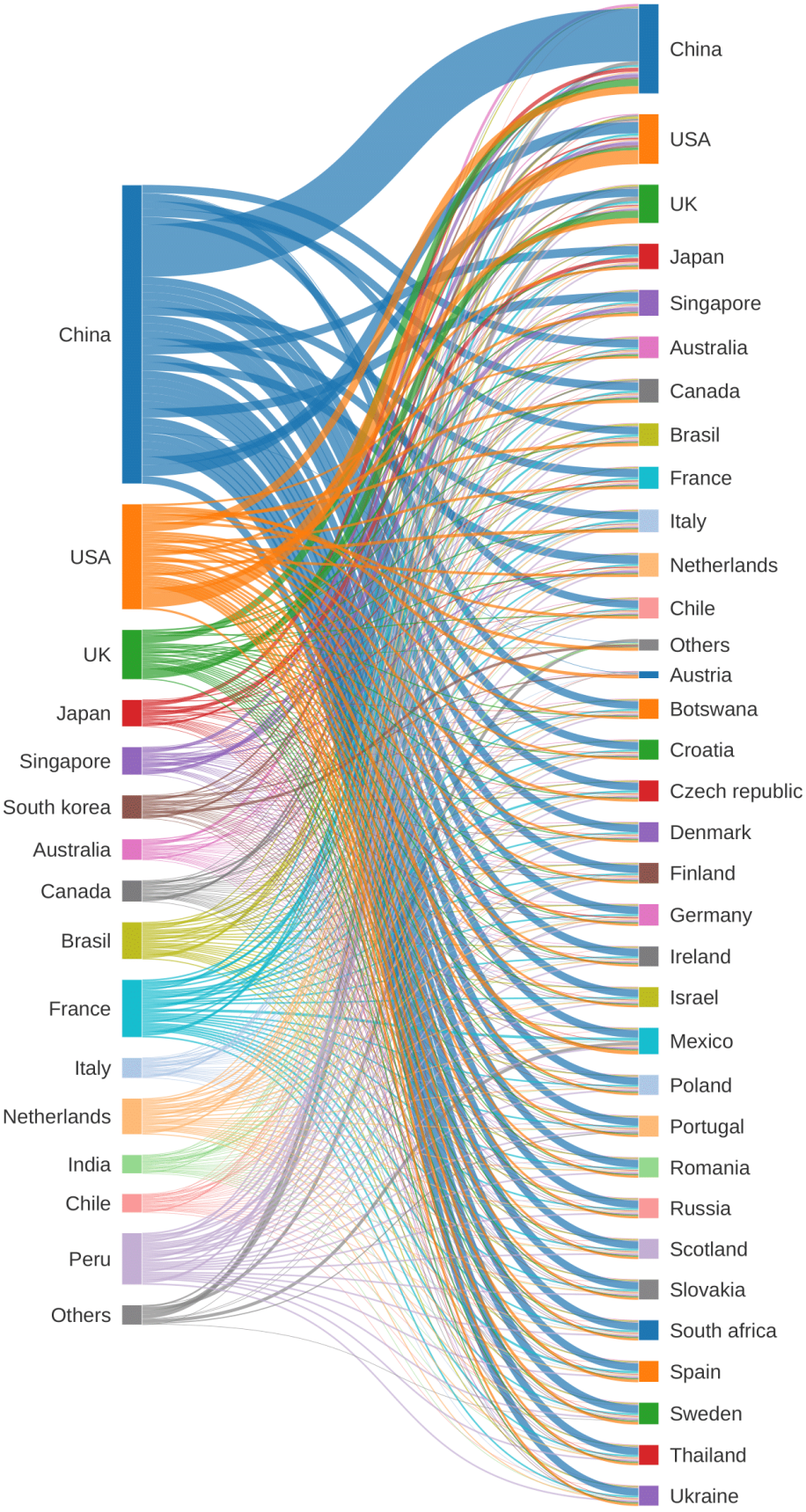

Figure 8 – The relation between author affiliations, country (left), and country applications studies (right). Source: the author.

Table 4 – Most used SVI service providers and the number of studies across different geographic scales. Studies may appear in multiple geo-scale categories.

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

this review, Google Street View (GSV), Tencent Street View (TSV), and Baidu Street View (BSV) are by far the most widely used. GSV offers the broadest international coverage, currently spanning more than 135 countries, whereas BSV and TSV provide similar functionality with a stronger focus on Chinese cities, covering approximately 652 and 296 cities, respectively.

Several regional alternatives also appear in the literature, including Naver Street View (NSV) and Kakao Street View (KSV) in South Korea, as well as CycloMedia in parts of Europe. In addition, community-driven platforms such as Mapillary and KartaView provide extensive image repositories that collectively contain approximately 1.5 billion images, although coverage and quality depend largely on user contributions. To better characterize data collection practices, we further classify studies according to their geographic scope: *city-scale* studies focusing on a single city, *country-scale* studies covering multiple cities within one nation, and *global-scale* studies spanning multiple countries. Table 4 summarizes the principal SVI services used throughout the literature. Because several studies employ more than one imagery provider and operate at multiple geographic scales, reported counts are non-exclusive.

3.2.5 Learning problems and tasks

Building on the learning taxonomy introduced in Chapter 2 (learning types, problems, and tasks), Table 5 summarizes the learning tasks identified across the selected studies, grouped by paradigm. Supervised learning dominates, led by classification (176 studies, 43.24%) and regression (134, 32.92%), reflecting the prevalence of both categorical labels and continuous perception scores in urban perception datasets. Unsupervised learning includes clustering (46, 11.30%) and dimensionality reduction (14, 3.44%) — the latter referring to techniques that compress high-dimensional SVI feature representations into lower-dimensional spaces for visualization or downstream modeling, complementing the clustering problem defined in Chapter 2. Self-supervised approaches — here encompassing generative models and VLMs, which learn representations without explicit human-provided labels — comprise multimodal language models (16, 3.93%), which began emerging in 2024, and generative models (6, 1.47%). Reinforcement learning, despite receiving dedicated

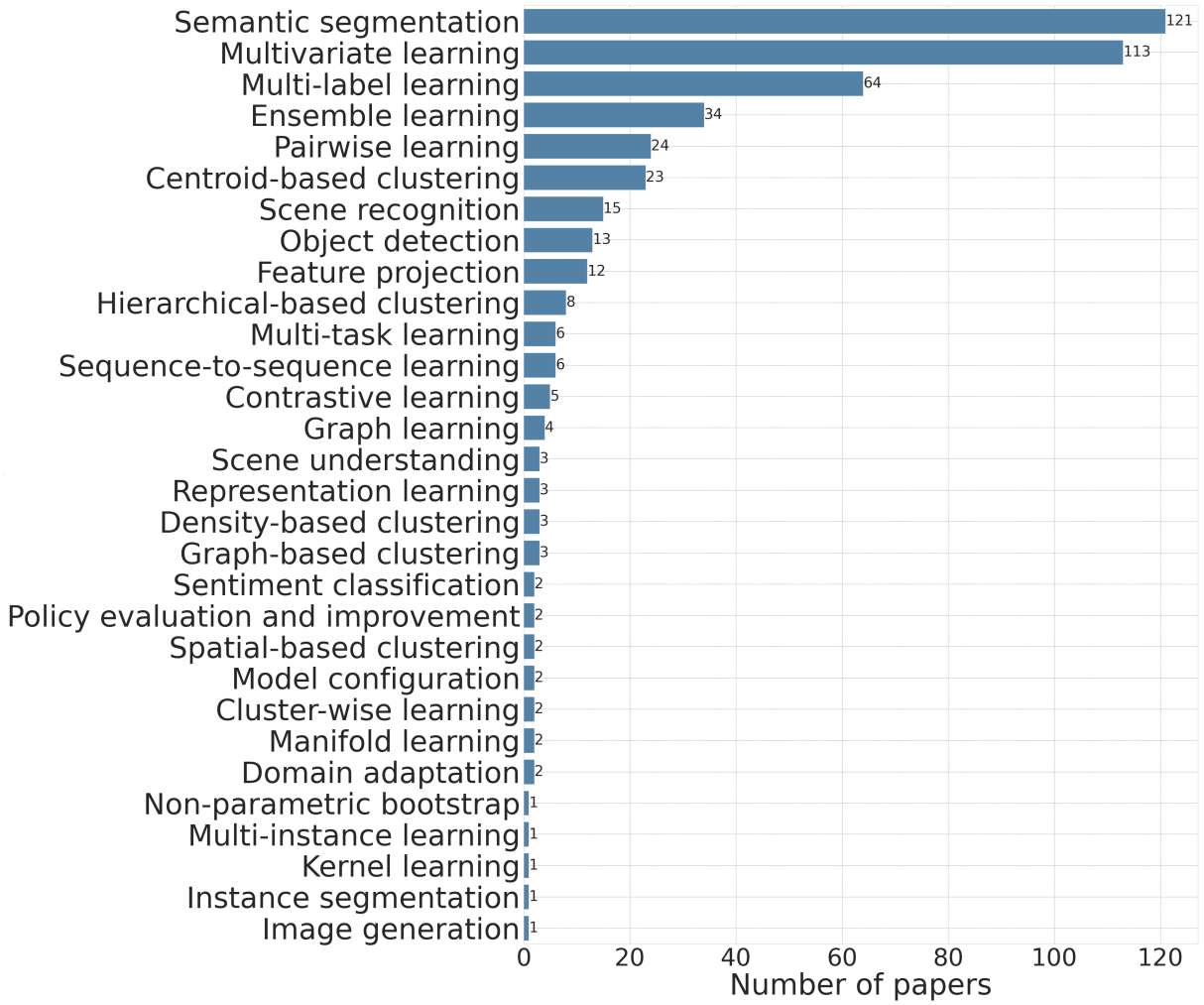

Figure 9 – The number of studies per task category shows that semantic segmentation and multivariate learning are the most common methods, followed by multi-label and ensemble learning. We note that some studies perform multiple tasks simultaneously, which places them in multiple categories.

coverage in Chapter 2 owing to its conceptual relevance, appears in only 2 studies (0.49%) within the SVI urban perception corpus, reflecting its limited practical adoption in this domain to date. Percentages are computed over the 407 total task occurrences, as many papers employed multiple tasks simultaneously.

Among specific tasks, semantic segmentation and multivariate learning are the most common (Figure 9), with many studies analyzing object-level pixel ratios and object importance. Rare tasks—appearing in at most one study—include *style transfer*, *policy learning*, *graph-based clustering*, *depth estimation*, and *kernel learning*.

3.2.6 Emerging Role of VLMs

An emerging trend across the reviewed literature is the increasing use of VLMs and LLMs in urban perception research. Beyond their application to rating, describing, and interpreting street-view scenes, recent studies have leveraged these models for tasks such

Table 5 – Learning type, learning problem, and number of studies.

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

as image editing, scenario generation, and counterfactual analysis, enabling researchers to explore how modifications to urban environments may influence human perceptions and experiences 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). This trend reflects a broader shift toward multimodal approaches that jointly integrate visual and linguistic information, allowing for richer representations of urban environments and more flexible frameworks for understanding, predicting, and shaping human perceptions of cities. This trajectory directly motivates the methodological choices adopted in the present thesis, which leverages VLMs both for perception assessment and for grounded counterfactual scene editing.

3.3 Final considerations

This chapter presented a systematic literature review of urban perception research using SVI, following the PRISMA protocol. Through a structured process of identification, screening, and eligibility filtering, an initial set of 2,307 papers was progressively narrowed to 137 studies. Applying forward and backward snowballing yielded 50 and 47 additional papers, respectively, resulting in a final corpus of **234 studies**. Descriptive analysis of this corpus revealed key patterns across publication trends, geographic distribution, SVI providers, and methodological approaches. Notably, research output has accelerated considerably since 2020, driven by the adoption of foundation models and multimodal learning techniques. An emerging trend across recent studies is the growing use of VLMs for urban perception tasks, signaling a broader shift toward richer, language-grounded representations of urban environments.

The review also revealed a set of persistent research gaps that motivate the contributions of this thesis:

- **Limited insight into visual factors**: the majority of studies concentrate on improving prediction accuracy while providing little explanation of which specific visual and semantic elements drive human safety judgments. *Addressed in Chapter 4.*
- **Static and indirect explanations**: when interpretability methods are applied, the resulting outputs — saliency maps or attribution scores computed over abstract feature representations — are often disconnected from the spatial layout of the scene itself and from planning-oriented scene modifications. *Addressed jointly in Chapters 4 and 5: the former grounds explanation in the spatial layout of the scene itself, the latter translates the resulting differences into human-readable recommendations.*
- **Unrealistic generative edits**: counterfactual and generative image-editing approaches tend to optimize edits to change a model's output without grounding transformations in differences observed between real urban scenes, limiting their credibility as evidence for design decisions. *Addressed in Chapters 6 and 7.*
- **Absence of interactive, human-in-the-loop tools**: no prior system integrates predictive modeling, grounded counterfactual generation, and interactive visual analytics within a single exploratory environment tailored to urban perception. *Addressed in Chapters 6 and 7.*

These four gaps map directly onto the three empirical contributions that follow. Chapter 4 takes up the first gap and begins closing the second by grounding explanation in the spatial layout of the scene itself, rather than in an abstracted feature vector. Chapter 5 completes the second gap by translating these spatially grounded differences into human-readable recommendations. Chapters 6 and 7 then close the third and fourth gaps together: UrbanCounterViz situates itself against the specific prior systems in visual analytics, computational perception modeling, and generative scene editing closest to it, grounds every scene edit in real, observed urban transformations rather than open-ended generation, and embeds the full pipeline in an interactive, human-in-the-loop environment.

4 Identifying Relevant Visual Elements in Urban Safety Perception

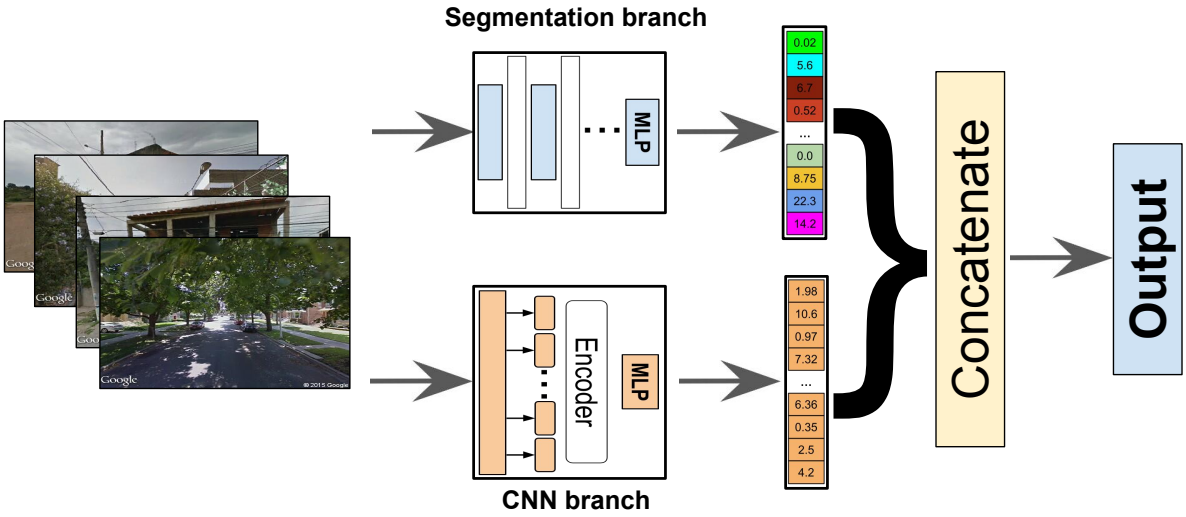

Figure 10 – The proposed UrbanFormer combines the OneFormer vectorized output with a modified ConvNext probability vector output to predict street human perception between safe and unsafe. Source: the author.

4.1 Introduction

This chapter presents the **first empirical contribution** of this thesis (Contribution 2, as enumerated in Chapter 1): an investigation of which visual and semantic elements most strongly influence human perceptions of urban safety. As discussed in Chapter 3, the majority of urban perception studies optimize for predictive accuracy — reporting how well a model classifies a street as safe or unsafe — while offering comparatively little insight into *why* a scene is perceived that way. Models trained directly on raw imagery, in particular, tend to behave as opaque predictors: they may achieve strong classification performance while leaving the visual reasoning behind their decisions largely inaccessible to the analyst.

This chapter addresses that gap directly, but the gap itself has two distinct faces in the prior literature. Models trained end-to-end on raw imagery achieve strong classification performance, yet are rarely subjected to a deep, element-level analysis of what the network's own attention is keyed on. Segmentation-based approaches, by contrast, do offer a form of interpretability — but typically only by treating the segmentation output as a static pixel-ratio feature vector and applying a post-hoc method such as LIME or SHAP to estimate the importance of each category within that vector. Neither route looks back

at the image itself: the first because it lacks a semantic vocabulary to describe what it attends to, the second because its explanation operates on an abstracted vector rather than on the spatial regions of the scene. This chapter proposes **UrbanFormer**, a classifier that fuses CNN visual features with an explicit semantic composition vector, paired with an explanation method that closes this remaining gap: rather than estimating feature importance over the pixel-ratio vector, it applies Grad-CAM to the CNN branch and intersects the resulting attention maps directly with the segmentation masks, measuring visual element relevance as spatial area overlap in the image itself. The work was published at the *IEEE/WIC/ACM International Conference on Web Intelligence* (WI-IAT '24) (Moreno-Vera et al., 2024). All materials can be found here[1].

Contributions. This chapter contributes: (i) UrbanFormer, a model that combines OneFormer-derived semantic composition vectors with CNN visual features, achieving the strongest reported binary classification performance among evaluated models and competitive performance on the finer-grained 10-label task; and (ii) an explanation methodology that crosses Grad-CAM activations with segmentation masks to quantify, element by element, which parts of a street scene are most associated with safe or unsafe perception — providing the visual vocabulary that Chapters 5 and 6 build on directly.

4.2 Related Works

Early approaches to computational urban perception relied on handcrafted image descriptors — including GIST features, DeCAF activations, and ImageNet-based representations — to model correlations between visual patterns and human judgments (Naik et al., 2014; Ordonez and Berg, 2014; Naik et al., 2016). These methods established the foundational pipelines still used today for collecting large-scale perceptual annotations through crowdsourcing, most notably the MIT Place Pulse dataset (Salesses et al., 2013), and demonstrated that perceptual attributes such as safety, beauty, and liveliness could be estimated from visual data at scale. However, the reliance on manually engineered features limited the representational capacity of these models and their ability to capture richer semantic and contextual relationships present in urban scenes.

The subsequent adoption of convolutional neural networks substantially improved predictive performance by learning hierarchical features directly from raw imagery, enabling models to capture finer-grained aspects of texture, layout, and spatial context (Dubey et al., 2016; Li et al., 2021; Zhang et al., 2021; Liu et al., 2022a; Zhou et al., 2019), and were subsequently applied to more targeted questions such as the influence of greenery and physical disorder on perceived safety (Li et al., 2015a; Porzi et al., 2015; Arietta et al., 2014; Tokuda et al., 2019; Lavi. et al., 2022). These studies established *what* can be

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

predicted from a SVI, but rarely examined which specific visual elements drove a given prediction.

More recently, segmentation-based approaches have improved interpretability by representing urban scenes in terms of semantically meaningful object categories, rather than raw pixels, and relating pixel-level scene composition to perceived safety (Zhang et al., 2018b; Min et al., 2020; Yao and Zhu, 2025; Zhou et al., 2025a; Liang et al., 2023; Zhu et al., 2024). These methods provide a more transparent basis for perception modeling than appearance-only CNNs, but typically use segmentation only as a preprocessing or visualization aid, rather than integrating it directly into the predictive model itself.

Differentiation. Prior work in this space follows one of two paths, and neither performs the specific kind of analysis this chapter introduces. Appearance-based CNNs achieve strong predictive performance but are seldom subjected to a deep, element-level explanation of what the network attends to. Segmentation-based approaches do provide a form of interpretability, but typically by applying a post-hoc method such as LIME or SHAP to the segmentation-derived pixel-ratio vector (Zhang et al., 2018b; Min et al., 2020) — estimating the importance of each semantic category as an abstract feature, without relating that importance back to where those categories actually appear in the image. UrbanFormer departs from both: it fuses CNN visual features with OneFormer's semantic composition vector for classification, and then explains its predictions by applying Grad-CAM directly to the CNN branch and intersecting the resulting attention maps with the segmentation masks themselves. Visual element relevance is thus measured as spatial area overlap between what the network attends to and where each element physically appears in the scene — a form of explanation grounded in the image, not in an abstracted feature vector. This joint design is what allows the chapter to answer not only *how safe* a scene is predicted to be, but *which elements* of the scene the prediction can be traced back to.

4.3 Methodology

The methodology is organized into three main steps. The first step involves **urban perception quantification**: an exploratory analysis of the Place Pulse 2.0 dataset to derive perceptual safety scores from pairwise image comparisons. The second step concerns the **UrbanFormer classifier**: a model that integrates a modified ConvNeXt Liu et al. (2022b) (hereafter CNN) with semantic features from the OneFormer Jain et al. (2023) segmentation model to predict perceived safety. The third step describes the **model explanation** technique, which applies Grad-CAM exclusively to the CNN branch in conjunction with segmentation masks to identify and quantify the contribution of specific visual elements to model predictions.

Unlike approaches that estimate feature importance by applying post-hoc methods

such as LIME or SHAP to the segmentation-derived pixel-ratio vector, this explanation method operates directly in image space: Grad-CAM attention maps from the CNN branch are intersected with the segmentation masks themselves, so that relevance is measured as spatial area overlap rather than as an importance score over an abstract feature vector.

4.3.1 Urban Perception Quantification

We build on the Place Pulse 2.0 dataset, introduced in Chapter 2. Of its six perceptual categories, this chapter focuses exclusively on *safety*, the attribute with the largest number of pairwise comparisons (368,926). Pairwise outcomes are converted into image-level Q-scores using the Strength of Schedule (SOS) algorithm (Park and Newman, 2005), defined in Chapter 2 (Equations 2.19a–2.19d) and scaled from 0 (very unsafe) to 10 (very safe). Q-scores are computed globally across all images.

4.3.2 Urban Elements location using semantic segmentation

We apply semantic segmentation to identify key visual elements within images, utilizing advanced models that classify each pixel into specific categories. This approach allows for the precise isolation and analysis of individual elements, enhancing our understanding of the spatial distribution and detailed characteristics of visual components. By leveraging semantic segmentation, we achieve a deeper, more structured insight into the contents of images, supporting a more comprehensive analysis of visual data. For this task, we select the most commonly used segmentation models in urban perception studies: PSPNet (Zhao et al., 2017), DeeplabV3+ (Chen, 2017), and SegFormerB5 (Xie et al., 2021) identified in Chapter 3. We also compare them with the OneFormer (Jain et al., 2023) segmentation model. When comparing SegFormer B5, OneFormer, PSPNet, and DeepLab v3 for semantic segmentation tasks, we note that OneFormer handles finer details, especially in street images; other models may not capture intricate boundaries as well. Moreover, OneFormer reported better segmentation task metrics than the others.

OneFormer is designed to handle both semantic and instance segmentation in a single, adaptable model. This allows OneFormer to capture nuanced spatial relationships and finer details with high precision, making it the preferred choice for detail-oriented segmentation applications. To qualitatively assess the segmentation performance, we randomly selected 100 segmented images produced by each method and manually inspected the resulting masks. This inspection focused on the correct delineation of object boundaries and the accurate segmentation of common urban elements such as sidewalks, roads, poles, vehicles, walls, and vegetation. Figure 11 presents representative segmentation masks generated by the evaluated models. Our manual inspection indicates that SegFormer, PSPNet, and DeepLabV3+ often fail to segment light poles, incorrectly merge walls, sidewalks, and gates into a single region, and produce incomplete or inaccurate vehicle

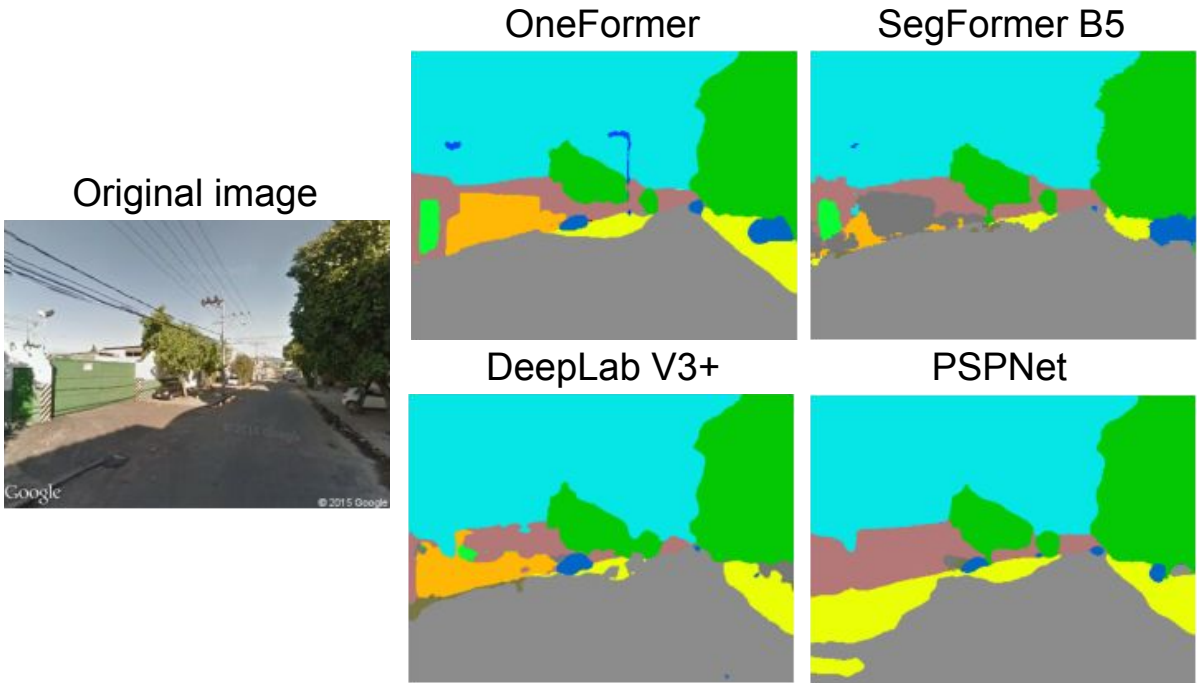

Figure 11 – Segmentation mask obtained from OneFormer, SegFormer B5, PSPNet, and DeepLabV3+. We note that some visual elements, such as a light pole, cars, gates, and walls, are incorrectly identified. Source: the author.

segmentations. In contrast, OneFormer consistently provides more detailed and accurate segmentations, particularly for small objects and complex scene boundaries. Although OneFormer requires greater computational resources and involves a more complex training and inference pipeline, its superior segmentation quality justifies these additional costs for applications requiring precise scene understanding.

4.3.3 UrbanFormer Classifier

We propose UrbanFormer, a classification model that combines visual and semantic representations of street-view scenes to predict perceived safety. The architecture integrates two pre-trained components: (i) a CNN backbone pre-trained on ImageNet, which extracts hierarchical spatial feature maps from raw RGB images and produces a compact feature vector via global average pooling; and (ii) the OneFormer segmentation model (Jain et al., 2023) pre-trained on ADE20K, which produces a pixel-ratio vector of size 150 corresponding to the 150 semantic categories defined in the ADE20K taxonomy.

The CNN component processes the input image through a series of convolutional stages, each progressively increasing representational depth while reducing spatial resolution, culminating in a high-level feature map that captures both local texture patterns and broader compositional structure. The OneFormer component outputs the proportion of each of the 150 semantic classes present in the image, encoding its semantic composition as a fixed-size feature vector. These two representations are fused by concatenation and

passed to a Multi-Layer Perceptron (MLP) head that performs classification, trained jointly with a cross-entropy loss. Figure 10 illustrates the UrbanFormer architecture.

The rationale for this fusion is that visual appearance features captured by CNN alone may not fully capture the semantic composition of a scene — i.e., *what objects are present* — which is independently informative for safety perception. By explicitly encoding the pixel-level presence of semantic categories such as trees, sidewalks, fences, and buildings, the OneFormer contribution allows the model to reason about scene content in a more structured and interpretable manner.

4.3.4 Model Explanation via Grad-CAM and Segmentation IoU

To identify which visual elements most influence safety predictions, we combine the Grad-CAM explanation method with the segmentation output of OneFormer. Grad-CAM is applied exclusively to the **CNN branch** of UrbanFormer, a design decision that follows directly from the structural properties of each component, as discussed below.

Why Grad-CAM is Applied Only to the CNN Branch.

Grad-CAM operates by computing the gradient of the target class score with respect to the activation maps of a chosen convolutional layer, then weighting those spatial maps to produce a localization heatmap over the input image. Its applicability therefore requires two conditions to hold simultaneously: (1) the target layer must produce *spatially structured* feature maps of shape $H \times W \times C$, preserving the correspondence between activation positions and image regions; and (2) the gradient of the classification output must flow *continuously and uninterrupted* through the computational graph back to that layer.

The CNN backbone satisfies both conditions. Its convolutional stages produce hierarchical spatial feature maps that retain spatial correspondence with the input image, and the gradient from the shared MLP classifier propagates cleanly back through global average pooling into these maps. Hooking into the last convolutional stage of CNN therefore yields a well-defined and spatially meaningful heatmap that reflects the model's joint decision — shaped by both the visual and semantic branches — for the target safety class.

The OneFormer branch, by contrast, does not satisfy these conditions in the context of the UrbanFormer pipeline. Although OneFormer internally produces rich spatial representations during segmentation — including pixel decoder feature maps and mask embeddings — what is actually extracted and passed to the MLP classifier is the pixel-ratio vector: a 150-dimensional histogram summarizing the proportion of each semantic class present in the image. This operation collapses all spatial structure, transforming a two-

dimensional segmentation output into a flat, orderless vector. No spatial correspondence between vector dimensions and image regions is preserved after this aggregation.

Furthermore, the most common implementation of this semantic vector extraction involves an arg max operation over class probability maps to produce a hard segmentation mask, followed by per-class pixel counting. The arg max function is non-differentiable: it does not admit a well-defined gradient and therefore creates a discontinuity in the computational graph. Even if OneFormer's internal convolutional layers are partially trainable, gradients from the binary classification loss cannot propagate through the arg max back to those layers. As a result, applying Grad-CAM to OneFormer's internal activations would not produce heatmaps that are meaningfully attributable to the classification decision; at best, such maps would reflect feature importance for the *segmentation* objective, not for the downstream binary safety prediction.

In summary, Grad-CAM is applicable to the CNN branch because it produces spatial feature maps through a differentiable pipeline connected end-to-end to the classifier. The OneFormer branch, by design, reduces its output to a non-spatial semantic descriptor via a non-differentiable operation before reaching the classifier, making spatial gradient-based attribution inapplicable to it.

Grad-CAM Formulation.

The Grad-CAM technique, formally introduced in Section 2.2, computes neuron importance weights for a target class c and feature map A^k as:

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

where α_k^c represents the neuron importance weight for feature map k and class c (consistent with the notation introduced in Chapter 2); $Z = u \times v$ is the spatial size of the feature map; A_{ij}^k denotes the activation value at position (i, j) in the last convolutional stage of CNN; and y^c is the predicted score for class c produced by the MLP classifier. The resulting localization map highlights the image regions in the CNN feature space most relevant to the model's prediction for the target safety class.

It is important to note that, because the gradient originates from the *fused* MLP classifier rather than from a *CNN-only* head, the resulting heatmap is influenced by the joint decision of both branches. In other words, the spatial attribution over the CNN maps implicitly encodes how the semantic composition captured by OneFormer contributed to shaping the model's grad-cam attention over image regions, even though the OneFormer branch itself cannot be directly visualized by Grad-CAM. Figure 12 illustrates this explanation pipeline.

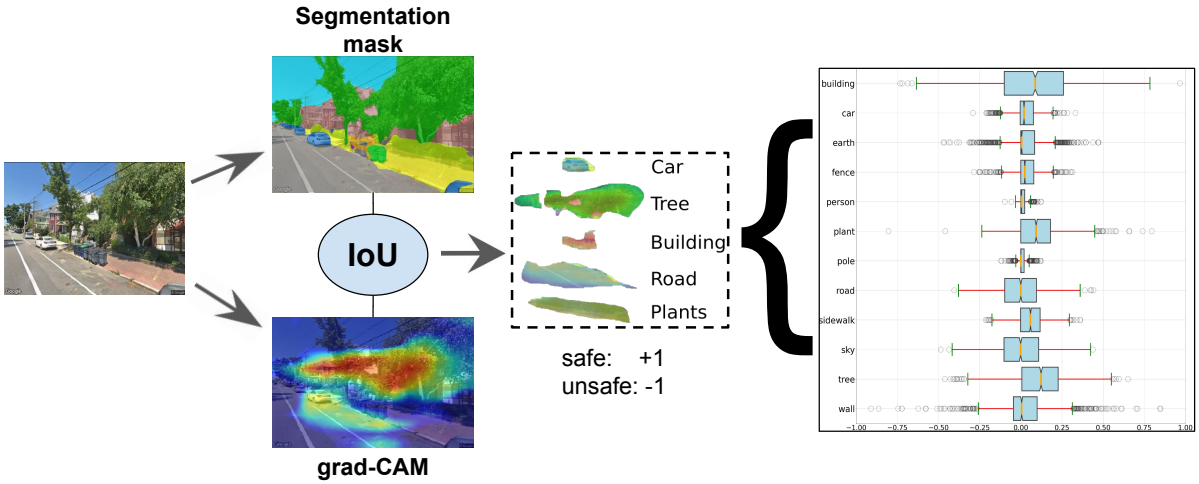

Figure 12 – Our explanation technique combines the segmentation mask and the Grad-CAM method to analyze the relationship between safety perception and the presence of visual elements. Source: the author.

Segmentation IoU for Element-Level Attribution.

To quantify the importance of individual semantic elements, we compute the **Intersection over Union (IoU)** between each segmentation mask produced by OneFormer and the Grad-CAM activation map generated by the CNN branch. A Grad-CAM threshold of 0.3 is applied to retain only sufficiently activated regions, following prior work (Vago et al., 2021; Li et al., 2025). The IoU score for each semantic category indicates how much of that object's area overlaps with the regions the model attends to when making its prediction. Positive weights are assigned to safe-image predictions and negative weights to unsafe predictions, enabling the analysis of which elements are associated with each class. As shown in Figure 13, the IoU metric is computed by comparing the binary mask derived from the grad-cam map with the corresponding ground-truth segmentation mask, where the overlap defines the intersection and the combined area defines the union.

4.4 Experiments

4.4.1 Dataset Exploration and Preparation

During dataset preparation, we identified 2,471 locations — defined as positions sharing identical latitude and longitude coordinates — assigned to more than one image ID. Representative examples include 130 repeated locations in Santiago (Chile), 126 in Berlin (Germany), and 112 in Montreal (Canada), down to 5 in Tel Aviv (Israel) and Seattle (USA). Mapping all duplicate IDs to unique locations reduces the effective dataset size from 111,390 to **108,820 images**.

We also observe a significant imbalance in the number of images across cities: some cities such as Amsterdam contribute as few as 622 images, while Atlanta contributes 3,965.

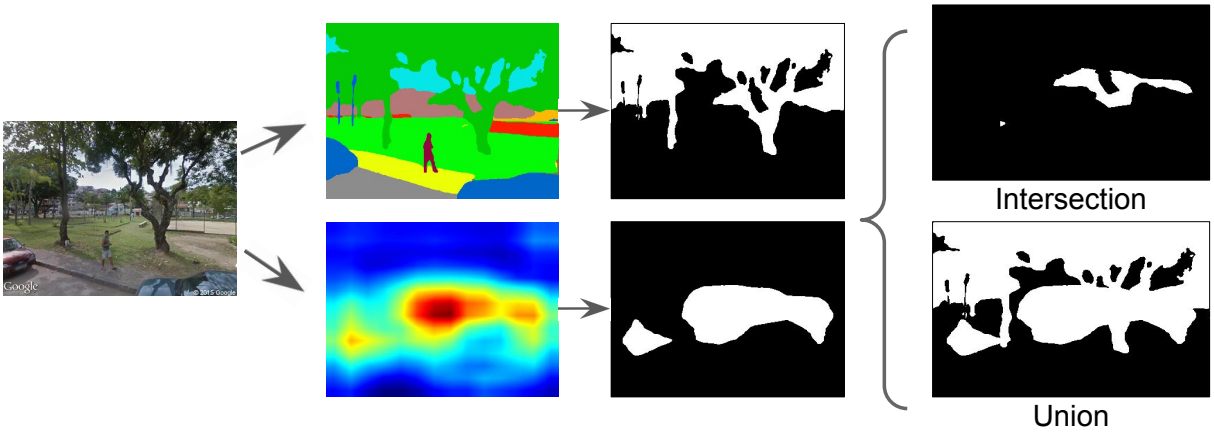

Figure 13 – Illustration of the Intersection over Union (IoU) computation for evaluating segmentation quality. The semantic segmentation map is used to extract the ground-truth mask of the target class (top), while the attention/saliency map is thresholded to produce the predicted mask (bottom). The IoU score is then computed as the ratio between the intersection (overlapping pixels) and the union (combined pixels) of the two binary masks. Source: the author.

Given this imbalance, all experiments are conducted using the full multi-city dataset.

As a final preprocessing step, we retained only images with at least five pairwise comparisons. This filtering criterion helps reduce the concentration of specific score values, such as 3.333 and 6.666, which arise from a limited number of comparisons and may introduce bias into the score distribution. Further details on this issue can be found in Moreno-Vera et al. (2021).

The dataset is partitioned into 75% for training and validation, and 25% for testing, using a class-stratified split so that the safe/unsafe proportion is preserved across partitions. To prevent data leakage, the collapsing of duplicate coordinate positions to unique image IDs described above is performed *before* partitioning, so that no physical location appears in both the training and test sets; the split is defined over these de-duplicated locations. Two classification tasks are defined based on the derived Q-scores: (i) **binary classification** (safe vs. unsafe, split at the median score); and (ii) **10-label classification**, where Q-score intervals [0–1), [1–2), . . ., [9–10] are mapped to integer labels 0 through 9. Model selection is carried out exclusively within the training and validation split via 5-fold stratified cross-validation (Section 4.4.2), leaving the test partition untouched during training and hyperparameter search.

4.4.2 Model Training and Performance

All model weights are initialized using the Xavier uniform criterion. Model selection is performed via grid search with 5-fold stratified cross-validation on the training and validation split to preserve class proportions across folds.

Table 6 – Accuracy comparison using binary classification (safe vs. unsafe)

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**

Tables 6 and 7 report the classification performance of UrbanFormer and prior methods on the binary and 10-label safety classification tasks, respectively. For binary classification, UrbanFormer-CNN-only already achieves competitive performance, while the complete UrbanFormer (CNN + OneFormer) reaches **75.22%** accuracy, outperforming all compared baselines. Although binary classification is useful for distinguishing broadly safe and unsafe scenes, it compresses the original perception scores into only two categories, discarding much of the underlying information.

The 10-label classification task provides a substantially finer-grained evaluation by partitioning the continuous safety perception scores into ten ordered intervals, preserving the score distribution and requiring the model to discriminate between subtle differences in perceived safety. Under this more informative setting, UrbanFormer (CNN + OneFormer) achieves **78.51%** accuracy, exceeding or matching all previous approaches, including segmentation-based fusion models. The strong performance on the 10-label task indicates that UrbanFormer effectively captures not only coarse safety distinctions but also the nuanced visual cues associated with gradual changes in urban perception.

Table 7 – Accuracy comparison using 10-label classification

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**

Overall, these results demonstrate that explicitly incorporating semantic scene composition through the OneFormer pixel-ratio vector consistently improves performance over appearance-only models. Furthermore, CNN's hierarchical convolutional representations provide a robust visual backbone, while the complementary semantic information enables

more accurate prediction across the full distribution of urban safety scores rather than only their binary extremes.

4.4.3 Model Explanations and Visual Element Importance

This analysis differs from the more common practice of estimating feature importance over a segmentation-derived pixel-ratio vector using LIME or SHAP; instead, it intersects the network's own Grad-CAM attention with the segmentation mask, directly in image space, to determine which visual elements the model actually attended to when making each prediction.

Figure 14(a) presents the three safest and three least safe images from the test set, alongside their corresponding OneFormer segmentation masks and Grad-CAM outputs computed over the CNN branch for the ground-truth label of each sample.

Applying the IoU analysis between Grad-CAM activations from the CNN branch and segmentation masks across all 108,820 images, we find that the most prevalent visual elements — trees, buildings, roads, sidewalks, walls, fences, ground, and sky — are present in activations from nearly 96% of the dataset images. Figure 14(b) reports the average impact of the 12 most prevalent visual elements, with positive values corresponding to safe predictions and negative values to unsafe ones.

The analysis reveals the following key findings regarding how visual elements relate to perceived safety:

- **Elements associated with safe perception**: Trees, grass, cars, and plants, present the strongest positive IoU overlap in safe images. Their consistent co-occurrence with model attention in safe predictions suggests that organized greenery and maintained infrastructure signal safety to both human observers and the model alike.
- **Elements associated with unsafe perception**: Walls, bare earth (dirt), fences, and trash receptacles exhibit higher IoU overlap in unsafe predictions, suggesting their association with environmental disorder and neglect — a result that aligns with the Broken Windows theory (Wilson and Kelling, 1982) discussed in Chapter 2.
- **High-prevalence, low-impact elements**: Elements such as roads, building, sidewalks, and sky are highly prevalent across all images but display a near-zero mean impact, indicating that their *spatial presence* — the mere area they occupy — is largely uninformative for distinguishing safe from unsafe scenes. We stress that this concerns the amount of sky visible, not the *condition* of the sky (weather, illumination, or time of day), which the static single-snapshot imagery and the pixel-ratio representation used here do not encode, and which prior work has shown to shape perception (Section 4.4.4).

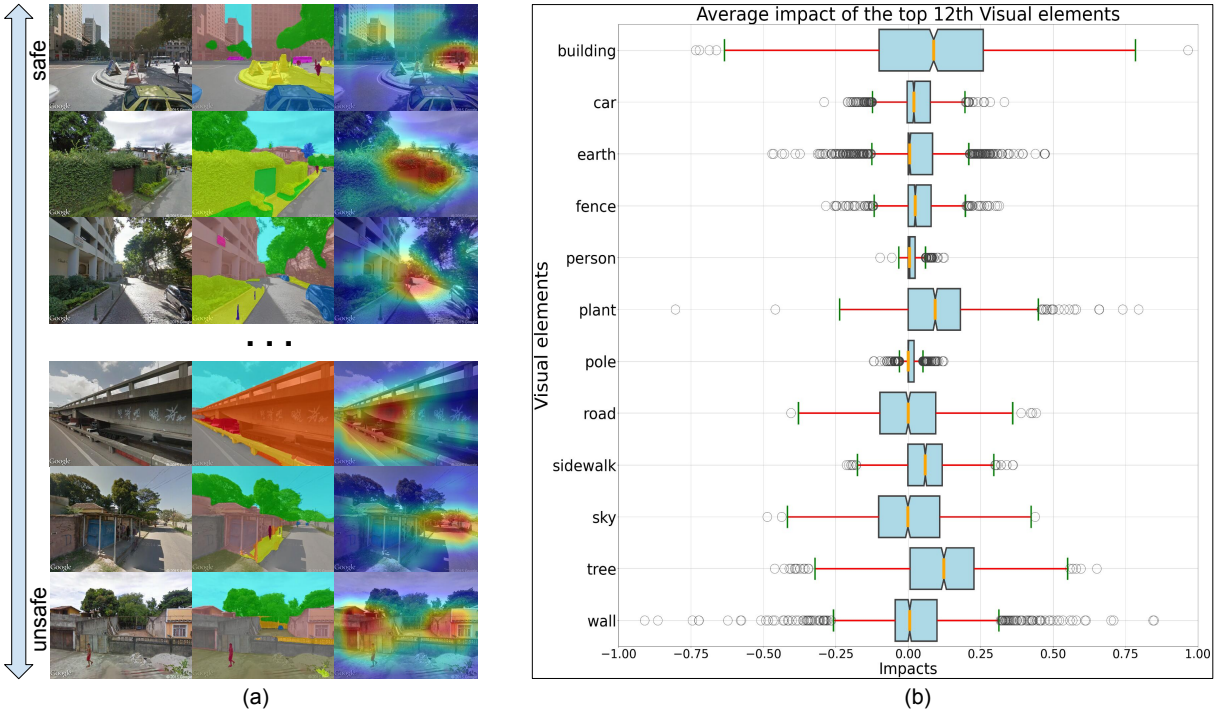

Figure 14 – We present the outputs of the Grad-CAM algorithm and the OneFormer output: (a) The top 3 safest and the top 3 less safe images, segmentation, and Grad-CAM outputs. (b) The average impact of the top 12 visual elements with presence in almost 90% of the 108,820 images from the dataset. We obtain this relevance by doing an IoU between the activations and the segmentation mask, where positive values correspond to safe predictions and negative values correspond to no safe predictions. Source: the author.

- **Rare elements**: Of the 150 visual categories identified by OneFormer, approximately 120 appear in fewer than 1% of images (e.g., flowers, pots, swimming pools) and contribute a near-zero average impact to the analysis.

Importantly, the analysis also suggests that the **absence** of certain elements (e.g., sidewalks, vegetation) may be as informative as their presence in characterizing unsafe scenes. Visual elements with high presence and a zero mean average impact indicate relevance to *both* safe and unsafe categories in equal measure; their absence, rather than their presence, may be the primary cue.

4.4.4 Limitations

Several limitations of this study should be noted. First, the Place Pulse 2.0 dataset does not support city-level analysis, since pairwise comparisons were assigned randomly across cities and the number of comparisons per city is insufficient for city-specific modeling. Second, the number of comparisons per image is highly unequal: most images were compared only three times, while a small number were compared up to 78 times, introducing variability in the reliability of individual Q-scores. Third, the class imbalance present after score-based

label assignment — across both binary and 10-label tasks — may bias model performance toward overrepresented classes, even after data augmentation.

Fourth, computational resource constraints limited batch sizes and the extent of hyperparameter and architecture search, potentially affecting the model's generalization capacity. Fifth, while Grad-CAM provides useful qualitative and quantitative insights into which image regions the model attends to, it reflects the model's learned attention rather than a causal account of human perception. Elements that appear frequently across all images (e.g., sky, roads) may inflate their IoU scores regardless of their true perceptual relevance.

A related caveat concerns the interpretation of high-prevalence, low-impact categories such as sky. Our analysis measures only the *spatial presence* of each element and treats the sky as a near-neutral category, but this reflects a property of the representation rather than of perception itself. Because each location is captured as a single static snapshot, the imagery holds weather, illumination, and time of day fixed, and the pixel-ratio vector encodes how much sky is visible but nothing about its appearance. Environmental criminology has repeatedly shown that these very factors matter: perceived safety and fear of crime vary with weather, season, and time of day, and homicide patterns in São Paulo have been shown to follow spatial–temporal and weather variations (Ceccato, 2005, 2012; Ceccato and Loukaitou-Sideris, 2022). The neutrality reported here should therefore be read narrowly — the *amount* of visible sky carries little discriminative signal in this dataset — and not as evidence that sky conditions are perceptually irrelevant. Capturing such effects would require temporally varying imagery and appearance-aware sky features, which we identify as an important direction for future work.

Sixth, the presence-based nature of the IoU relevance metric systematically favors large, ubiquitous elements over small but salient ones. Because relevance is measured as area overlap and then averaged across the dataset, an element that occupies only about 1% of an image (e.g., graffiti or a damaged sign) contributes a near-zero mean impact even when it strongly attracts the model's attention wherever it does appear, whereas pervasive elements such as trees accumulate high overlap largely by virtue of their size and frequency. The metric therefore rewards prevalence rather than per-object saliency, and small disorder cues may be under-credited relative to their true perceptual weight. This limitation, together with possible corrections, is discussed in more depth as a cross-cutting issue in Chapter 8.

Finally, the scope of gradient-based explanation in UrbanFormer is inherently constrained to the CNN branch, as discussed in Section 4.3.4. The OneFormer branch contributes to the classification decision through its pixel-ratio vector, but this contribution cannot be spatially visualized via Grad-CAM due to the non-spatial and non-differentiable nature of the semantic vector extraction process. Future work should explore complementary

attribution methods — such as Integrated Gradients or SHAP — applied to the semantic vector dimensions to disentangle the class-level contributions of individual semantic categories from those captured by the spatial heatmaps. Mechanisms that explicitly disentangle occurrence frequency from semantic importance also remain an open direction.

4.5 Final Considerations

This chapter presented UrbanFormer, a model fused with semantic segmentation features for urban safety perception classification. The architecture combines the hierarchical convolutional representations of CNN with the pixel-ratio vectors produced by OneFormer, achieving competitive performance on both binary and 10-label classification tasks on the Place Pulse 2.0 dataset, and outperforming prior segmentation-based and CNN-based approaches.

More importantly, the Grad-CAM and IoU-based explanation analysis — applied to the CNN branch and crossed with OneFormer segmentation masks — provides interpretable evidence of which visual elements most strongly shape perceived safety in SVI. Trees, sidewalks, and organized vegetation are consistently associated with safe perceptions, while walls, bare ground, and trash receptacles are linked to unsafe ones. These findings provide an empirically grounded basis for urban design recommendations and link data-driven model behavior to the broader theoretical literature on urban disorder.

These findings motivate what follows, but knowing that an element matters is not yet knowing what a street would need in order to feel different: Grad-CAM and IoU show where a model looks, not how a scene could plausibly change. Chapter 5 takes up exactly this question, translating counterfactual difference vectors — which visual elements should change — into plain-language recommendations that may support early-stage reasoning by urban planners. To do so, it also introduces UrbanPD4k, a dataset of 3,659 Rio de Janeiro SVIs with pixel-level annotations of 13 physical disorder elements not covered by standard segmentation taxonomies. Chapter 6 then goes further still, building on both the visual vocabulary established here and the counterfactual reasoning of Chapter 5 to make such changes visible and inspectable, rather than only described.

5 Counterfactual Explanations and Human-Readable Interpretations of Urban Safety Perception

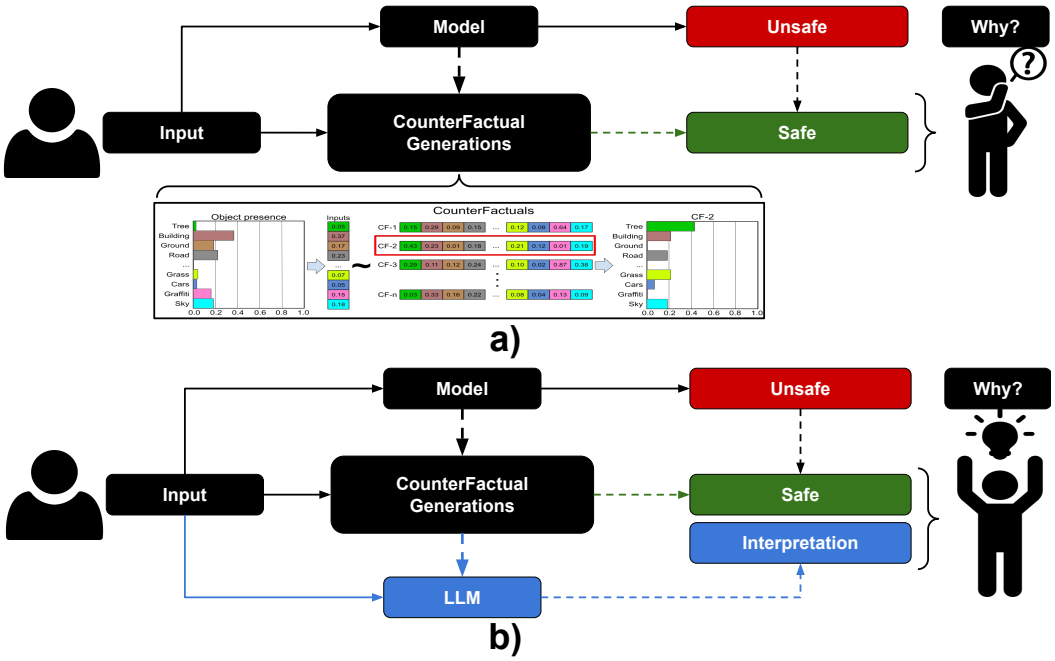

Figure 15 – Overview of UrbanCFX (Urban CounterFactual eXplanations) for safety prediction and interpretation. The model initially predicts an unsafe outcome. (a) Counterfactuals suggest input changes that yield a safer prediction, though they remain difficult to interpret. (b) An LLM generates human-readable explanations of these counterfactuals.

5.1 Introduction

Chapter 4 established which visual elements are consistently associated with perceived street safety and quantified their relevance through spatial model attention. That analysis, however, stops at diagnosis: knowing that walls, bare earth, or trash receptacles matter does not indicate what should change about them, nor does it produce recommendations that planners or policymakers can act on directly. The explanations produced by Grad-CAM and IoU are numeric and spatial—overlap scores tied to image regions—and interpreting them still requires a human to translate heatmaps and percentages into concrete interventions. Addressing this limitation constitutes the **second empirical contribution** of this thesis (Contribution 3, introduced in Chapter 1). Specifically, this chapter identi-

fies scene changes that could plausibly improve perceived safety and expresses them as natural-language recommendations that non-specialists can readily understand.

Two limitations of the vocabulary used in Chapter 4 motivate the first step of this chapter. Standard segmentation taxonomies such as ADE20K describe a street's structural layout — walls, sidewalks, roads — but do not distinguish a maintained wall from a graffiti-covered one, or an intact sidewalk from a broken one. Yet it is precisely this finer distinction, well documented in the urban perception literature since the "Broken Windows" theory (Wilson and Kelling, 1982), that most directly signals disorder to a human observer (Nasar, 1998; Ulrich, 1979; Schroeder and Anderson, 1984). This chapter therefore extends the semantic vocabulary established previously with a set of manually annotated physical disorder elements — graffiti, damaged walls, overhead cables, broken sidewalks, among others — capturing the specific cues most associated with perceived neglect.

With this richer vocabulary in place, the chapter turns to its central problem: producing an explanation a person, not only a model, can use. Counterfactual analysis identifies which minimal combination of element changes would shift a scene's classification from unsafe to safe — but the direct output of that analysis is a numeric difference vector, no more immediately usable than the heatmaps of the previous chapter. The core methodological contribution of this chapter closes that gap with a LLM: rather than presenting an analyst with a list of features and signed values, the pipeline developed here, **UrbanCFX**, converts each counterfactual difference vector into a plain-language sentence describing what changed and why it matters. This is the throughline that distinguishes this chapter from the one before it: Chapter 4 answers *which* elements matter and shows that answer as a chart of attributions; this chapter answers *what would need to change* and expresses that answer as text a person can read directly, without first interpreting a vector, a score, or a heatmap themselves. This work was published at the *IEEE International Conference on Big Data* (BigData '25) (Moreno-Vera et al., 2025). All materials can be found here[1].

Contributions. This chapter contributes: (i) **UrbanPD4k**, a dataset of 3,659 Rio de Janeiro SVIs with pixel-level manual annotations of 13 physical disorder elements not covered by standard segmentation taxonomies; (ii) an empirical comparison showing that these disorder elements carry stronger predictive and counterfactual signal for perceived unsafety than general segmentation categories do; and (iii) **UrbanCFX**, a pipeline that generates counterfactual difference vectors and, critically, translates them into human-readable natural language using a LLM — turning a numeric, model-facing explanation into a text-based, human-facing one. This last contribution is the one that most directly advances the thesis as a whole: it marks the first point in the dissertation's progression

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

where a machine-derived signal is expressed in the same language a planner already reasons in, rather than in the language of the model that produced it.

5.2 Related Works

Urban street disorder elements have long been recognized as influential in how people perceive and evaluate urban environments. Under the "Broken Windows" theory (Wilson and Kelling, 1982), visible signs of disorder — graffiti, litter, damaged infrastructure — negatively affect perceptions of safety and social stability, and studies in environmental psychology have shown that observers interpret such cues as signals of neglect, shaping feelings of comfort and trust (Nasar, 1998; Ulrich, 1979). Nasar and Jones (Nasar and Jones, 1997) found that neighborhoods exhibiting physical disorder were rated significantly lower in perceived safety and pleasantness. More recent computational work has correlated specific disorder elements, particularly graffiti, with socioeconomic indicators such as the Human Development Index (Diniz and Stafford, 2021; Alzate et al., 2021; Lavi. et al., 2022; Tokuda et al., 2019). Despite this evidence, disorder-specific categories remain largely absent from the segmentation taxonomies used in mainstream urban perception modeling, limiting their availability to perception-centered analysis — a gap this chapter's annotation effort addresses directly.

On the explanation side, Chapter 4 already discussed how segmentation-based perception models are commonly interpreted by applying LIME or SHAP to a pixel-ratio or presence-based feature vector (Zhang et al., 2018b; Min et al., 2020; Ma et al., 2025; Wu et al., 2025), producing an importance score per semantic category. Counterfactual explanation methods go a step further, identifying the minimal set of feature changes that would flip a model's prediction rather than merely ranking feature importance (Mothilal et al., 2019), and have been applied in the broader explainable-AI literature to generate such difference vectors for image-based classifiers (Gomez et al., 2020; Augustin et al., 2022). In every one of these cases, however, the output delivered to the analyst remains a vector, a table, or a heatmap — a representation that still requires the reader to translate signed feature values into an actionable judgment. Separately, other work has used language models to describe urban scenes directly, through image captioning or multimodal description (Ma, 2023; Ma and Wu, 2023), but these approaches summarize *what is present* in a scene rather than *what should change* to alter how it is perceived.

Differentiation. UrbanCFX occupies the space these two lines of work leave open. It does not stop at a counterfactual difference vector, as prior counterfactual methods do, nor does it use a language model merely to describe a scene, as captioning-based approaches do. Instead, it uses the language model as a translation layer positioned directly over the counterfactual output: structured prompting grounds the LLM in the exact difference vector produced for a given image, and its role is specifically to render that vector —

numeric, structured, and otherwise legible only to someone versed in the model's feature space — into a natural-language sentence describing what changed and what that change means for perceived safety. Where Chapter 4 produced an explanation that remains a chart, this chapter produces one that reads as a recommendation.

5.3 Methodology

The methodology is organized into four stages. The first stage involves **image segmentation and manual annotation**: applying OneFormer to extract standard ADE20K semantic features and enriching them with human-annotated physical disorder elements. The second stage concerns **dataset and city selection**: motivating the choice of Rio de Janeiro as the target city within Place Pulse 2.0 and describing the image subset used for all experiments. The third stage performs **counterfactual explanation**: generating minimal feature-space edits that shift classification from unsafe to safe, under two complementary feature representations. The fourth and central stage introduces **LLM-driven human-readable interpretation**: mapping counterfactual difference vectors to natural language through structured prompting of a LLM.

Table 8 – ADE20K classes and new group-categories (visual elements)

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)

5.3.1 Image Segmentation and Manual Annotations

As in Chapter 4, semantic segmentation is performed using OneFormer pre-trained on ADE20K (Jain et al., 2023), producing a pixel-ratio vector of 150 semantic classes for each image. For this study, these 150 classes are regrouped into eight higher-level **visual element** categories — **Sky**, **Human**, **Construction**, **Floor**, **Vegetation**, **City**

elements, **Terrain vehicles**, and **Body of water** — to reduce dimensionality while preserving semantic interpretability. Table 8 summarizes the regrouped categories.

Crucially, this standard vocabulary is extended with manual pixel-level annotations of **13 physical disorder elements**, identified through psychological and sociological literature on urban environments (Wilson and Kelling, 1982; Lynch, 1984; Schroeder and Anderson, 1984; Nasar, 1998). These elements include **overhead cables, graffiti, broken pavement, garbage bags, trash cans, damaged walls, broken windows, broken sidewalks, deteriorated traffic signs, police vehicles, homeless individuals, and related indicators of physical decay and social vulnerability**.

For this step, two trained annotators independently labeled the presence at pixel-level of each cue in all n_i images, followed by reconciliation to resolve disagreements ($\kappa = 0.81$ inter-rater agreement). Both annotators are PhD students — including the author of this thesis — with prior research experience in urban perception and hands-on familiarity with the Place Pulse dataset, and both have lived in Rio de Janeiro for more than two years. This shared background provided consistent, locally grounded criteria for identifying context-specific disorder cues (such as densely bundled overhead cables or informal-settlement wall conditions) that a non-local annotator might overlook or misclassify. These annotations allow fine-grained analyses of how visible disorder correlates with perceived safety, capturing scene qualities that no pre-existing segmentation taxonomy covers. Figure 16 illustrates the process of merging ADE20K segmentation masks with the new disorder element annotations.

The motivation for this annotation effort follows directly from Chapter 4: elements such as bare earth, trash receptacles, and walls were associated with unsafe perceptions, but standard segmentation models do not distinguish between a well-maintained wall and one covered in graffiti, or between an intact sidewalk and a broken one. The 13 disorder annotations address precisely this representational gap.

5.3.2 Dataset and City Selection

The dataset used in this study is drawn from Place Pulse 2.0, restricted to images from the city of Rio de Janeiro, yielding 3,659 street-level scenes with their corresponding pairwise safety comparison outcomes. The choice of Rio de Janeiro is not arbitrary: among all 56 cities included in Place Pulse 2.0, Rio de Janeiro records the lowest average perceived safety score (Dubey et al., 2016; Moreno-Vera et al., 2021), making it the most perceived safety-critical urban context in the dataset and, consequently, the most informative for studying the visual elements associated with unsafe perceptions. Figure 17 illustrates the spatial distribution of perceived safety across the city, together with the geographic coverage of the sampled SVIs used in this study.

This low perceived safety profile is strongly linked to the high prevalence of physical

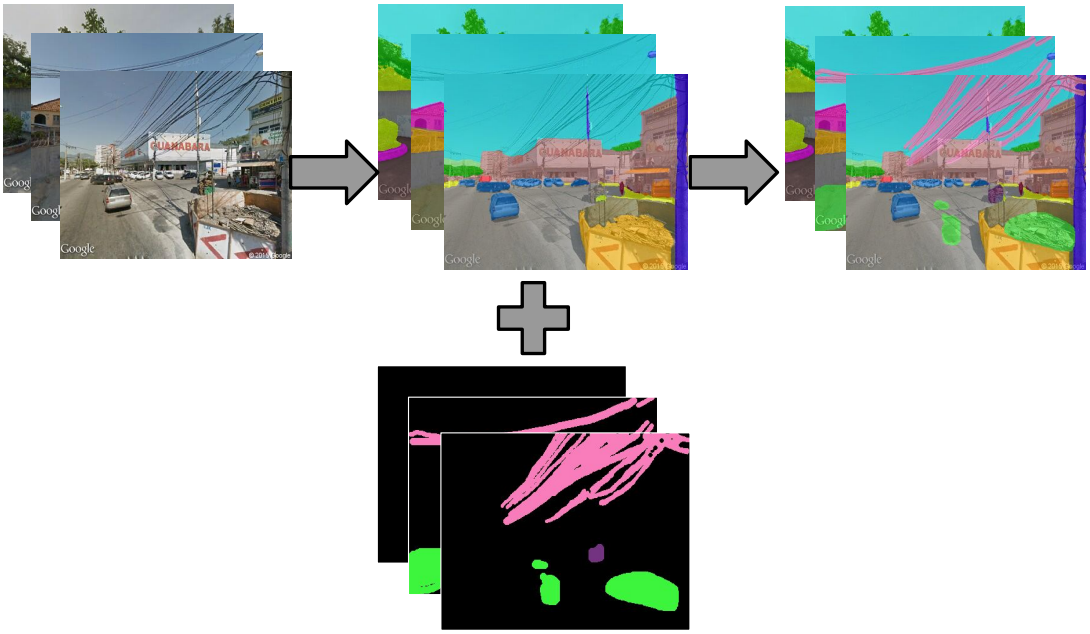

Figure 16 – To generate a complete segmentation mask per sample, we perform a replacement at pixel level. Source: the author.

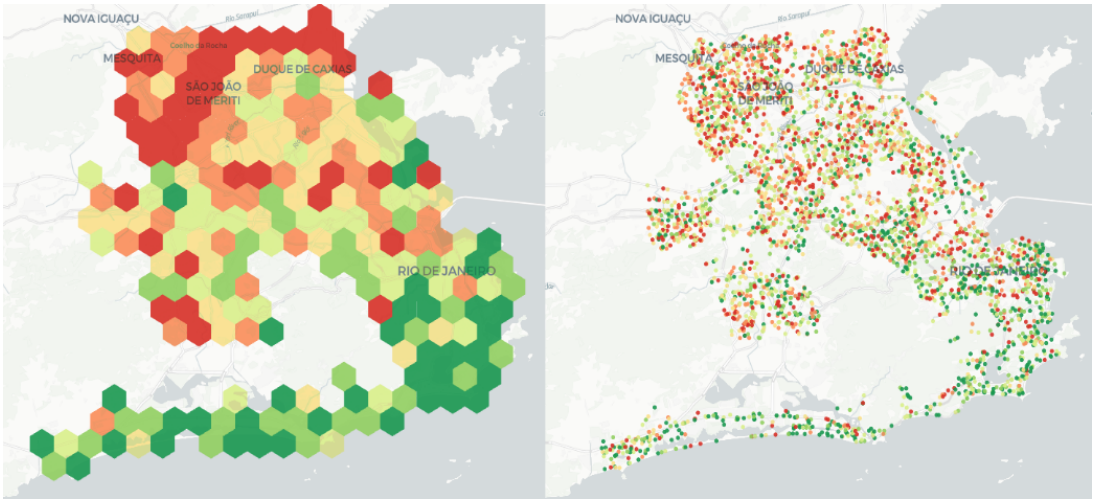

Figure 17 – Left: Hexagonal aggregation of the median perceived safety score (Q-score) computed from Place Pulse 2.0 street-view images, where green indicates higher perceived safety and red indicates lower perceived safety. Right: Geographic distribution of the 3,659 sampled street-view images, colored according to their binary safety label (green = safe, red = unsafe), illustrating the spatial coverage of the dataset across the city. Source: the author.

disorder elements in Rio's streetscapes. The city's urban fabric is marked by stark socio-spatial contrasts (Diniz and Stafford, 2021): affluent neighborhoods such as Copacabana and Ipanema coexist alongside extensive informal settlements (favelas), where approximately 23% of the population lives in conditions of deficient infrastructure and limited public services (Liu et al., 2025). These informal areas are characterized by visible signs of physical disorder—overhead cables running between buildings in dense webs, deteriorating walls, graffiti, broken sidewalks, garbage accumulation, and damaged infrastructure—that

are substantially more prevalent than in the formal city and that serve as the primary visual cues through which residents and observers form safety perceptions (Remigio et al., 2019). This concentration of disorder elements, precisely those captured by the 13 manual annotations introduced in this chapter, makes Rio de Janeiro a uniquely rich and representative context for studying their impact on safety perception.

From a methodological standpoint, focusing on a single city also ensures that all images share a coherent visual grammar—the same mix of architectural styles, infrastructure typologies, and environmental conditions—avoiding the inter-city confounds inherent in the broader multi-city dataset analyzed in Chapter 4. Perceptual safety scores for all images are computed using the Strength of Schedule (SOS) algorithm, as described in Section 2.4.2, and binary classification labels are assigned based on the median Q-score.

5.3.3 Counterfactual Explanations

Counterfactual explanations, formally introduced in Chapter 2 (Section 2.2), seek the minimal change to an input that would alter the model's prediction. Here, counterfactuals are generated in **feature space** rather than pixel space: the input x is a vector of visual element descriptors, and the counterfactual x' specifies which elements must change — in presence or quantity — to produce a safe classification. Two feature representations are evaluated throughout the chapter.

Binary presence/absence.

Each feature $x_i \in \{0, 1\}$ indicates whether a visual element is present. The counterfactual objective minimizes the number of flipped indicators:

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

where τ is the safety perception classification threshold and $\mathcal{F}$ is the set of feasible edits. The difference $\Delta x_i = x'_i - x_i \in \{-1, 0, 1\}$ indicates whether element i was removed, unchanged, or added.

Pixel ratios.

Each feature $x_i \in [0, 100]$ encodes the proportion of the image area occupied by a visual element. The counterfactual objective seeks the closest vector that achieves the target class:

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

where λ balances proximity and prediction shift. The real-valued difference $\Delta x_i = x'_i - x_i$ captures the magnitude of coverage change for each element, enabling more nuanced

counterfactual reasoning than binary flips: the change from 10% to 25% tree coverage, for instance, is treated as a gradual intervention rather than a simple presence toggle.

5.3.4 Human-Readable Interpretations via LLM

The central contribution of this chapter is the transformation of counterfactual difference vectors into natural language through a LLM. Once the counterfactual x' is obtained, the element-level differences $\Delta x = x' - x$ are structured into two sets — elements whose presence or coverage increased, and elements whose presence or coverage decreased — and passed to an LLM through a structured prompt. The LLM then produces a plain-language recommendation describing what visual changes would improve the scene's perceived safety. Prompt 1 illustrates the template used to query the model, where `added` and `removed` are dynamically filled with the counterfactual insights.

```
[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 – LLM prompt used for street safety perception evaluation

Formally, this mapping is expressed as:

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

For the binary case, where differences are discrete, the explanation is derived from the added and removed element sets:

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

For the pixel-ratio case, a threshold ϵ filters out negligible fluctuations before interpretation:

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

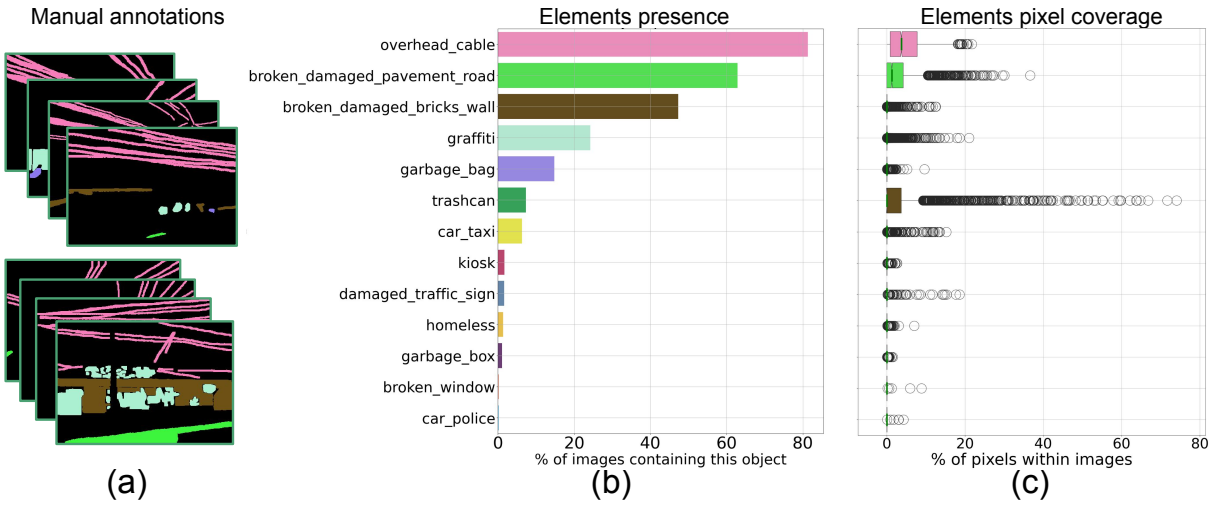

Figure 18 – Overview of the annotation statistics for the UrbanPD4k dataset. (a) Pixel-level annotation mask samples (b) Presence of annotated elements, indicating the frequency of each class in the dataset. (c) Pixel coverage of elements, showing the proportion of the image area occupied by each annotated class. Together, these visualizations highlight the diversity and balance of the annotated content within UrbanPD4k. Source: the author.

ensuring that only semantically meaningful changes reach the model — for example, graffiti coverage dropping from 15% to 2%, or tree coverage rising from 10% to 25% — rather than numerical noise.

Prior work has compared and evaluated the performance of LLMs in generating grounded textual explanations (Khan et al., 2024; Zytek et al., 2024), finding that models in the GPT-4 family achieve the strongest performance. Based on these findings, we use GPT-4o mini, which provides a favorable balance between explanation quality and computational cost. Prompts were structured following established guidelines for plain-language explanation generation (Wang et al., 2024), instructing the model to describe the visual changes in terms accessible to urban planners without attributing false causality or introducing information absent from the difference vector. This grounding condition is essential: the LLM is not asked to reason freely about urban safety in general but to translate a specific, model-derived intervention into natural language.

5.4 Experiments

5.4.1 UrbanPD4k Dataset and Annotation Statistics

We introduce the UrbanPD4k dataset, which comprises 3,659 SVIs from Rio de Janeiro with pixel-level annotations across 13 physical disorder categories. Figure 18 summarizes the annotation statistics through three complementary visualizations. Subfigure (a) presents representative examples of the generated segmentation masks across the annotated categories, illustrating the extent of human involvement in the labeling process.

Subfigure (b) reports the occurrence frequency of each annotated class, providing an overview of how often different urban physical disorder elements—such as cables, trash cans, garbage, and graffiti—appear throughout the dataset. Subfigure (c) presents the pixel coverage of each class, reflecting their relative spatial occupancy within the scenes. Together, these visualizations show that large structural classes, such as buildings and roads, account for the majority of the image area, whereas smaller yet semantically significant elements—including damaged walls, broken windows, homeless individuals, and deteriorated traffic signs—occur less frequently but substantially enrich the visual and contextual diversity of urban environments. Finally, the dataset is split into 75% for training and validation and 25% for testing across all experiments, using a class-stratified partition that preserves the safe/unsafe proportion and keeps the test set fully held out during model fitting, consistent with the leakage-safe protocol described in Chapter 4.

5.4.2 Classification Performance and Disorder Impact

Four feature configurations are evaluated under both binary and pixel-ratio settings: (i) ADE20K classes alone (150 features); (ii) ADE20K combined with disorder annotations (163 features); (iii) grouped visual elements alone (8 features); and (iv) grouped visual elements combined with disorder annotations (21 features).

A Random Forest classifier trained on the feature vectors described above serves as the underlying prediction model. Unlike the deep neural network used in Chapter 4, the Random Forest operates on named, interpretable features, making counterfactual differences directly mappable to identifiable visual elements and thus suitable for language-based interpretation. We also evaluated alternative classification models, including linear methods, bagging-based ensembles, and boosting algorithms. However, these approaches consistently achieved inferior predictive performance on our dataset, making the Random Forest the most suitable choice for balancing classification accuracy and interpretability.

Table 9 reports the classification results. Across both feature representations, incorporating disorder elements consistently improves performance. In the binary setting, adding disorder annotations to the ADE20K vocabulary improves F1 scores for unsafe detection by 2–5%; in the pixel-ratio setting, the improvement reaches 5–7%. The best overall configuration — grouped visual elements combined with disorder annotations in the pixel-ratio setting — achieves F1 scores of 0.75 (not safe) and 0.72 (safe). These results confirm that disorder-related elements carry meaningful predictive signal beyond what general segmentation categories provide, and that pixel-ratio representations capture finer-grained perceptual variation than binary presence indicators.

To further understand the contribution of disorder features, SHAP explanations (Lundberg and Lee, 2017) are applied to both representations. In the pixel-ratio setting, SHAP assigns positive and negative contributions to all elements proportionally to

Table 9 – Classification results using binary and pixel ratios configurations

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**

their marginal influence: disorder elements consistently receive negative SHAP values, reflecting their association with lower perceived safety, while vegetation and well-maintained construction elements contribute positively.

Critically, the magnitude of these contributions is asymmetric: disorder elements drive unsafe predictions more strongly than their absence, or the presence of orderly elements, drives safe ones — disorder functions primarily as a signal of danger rather than its absence functioning as an equally strong signal of safety perception. This is consistent with the Broken Windows framing that motivates the UrbanPD4k annotation effort: visible neglect appears to carry more perceptual weight than visible order.

In the binary setting, absent elements receive near-zero SHAP attribution, limiting the explanatory coverage. **This asymmetry reinforces the pixel-ratio representation as the stronger basis for both classification and explanation**.

This SHAP-based analysis, like the segmentation-based explanations discussed in Chapter 4, remains an aggregate, feature-level account: it tells us *which elements matter* across the dataset, not what a specific scene needs. The following subsection addresses that instance-level gap directly.

5.4.3 Counterfactuals and LLM Interpretations

Counterfactual analysis identifies the elements most frequently modified when shifting predictions from unsafe to safe. Among disorder elements, overhead cables, graffiti, damaged walls, garbage bags, and broken sidewalks exhibit the highest modification frequency across the 100 generated counterfactuals per image using the DiCE library (Mothilal et al., 2019). Among general elements, vegetation and human presence are also commonly

involved. The frequency with which an element appears in minimal transformations is interpreted as a proxy for its counterfactual relevance: the elements the classifier most relies on to predict unsafety are precisely those that most often need to change to produce a safe prediction.

Table 10 – LLM human-readable interpretations for Counterfactuals

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.

Table 10 presents LLM-generated interpretations for three representative unsafe images, corresponding to high, medium, and low perception gaps (Δsafety). For the image with the largest shift (Δsafety = 0.45), where graffiti, garbage, and a damaged sidewalk are removed and trees, walls, and fences added, the LLM produces: *"It's better to remove graffiti from walls, repair brick walls, and avoid overhead cables."* For a moderately unsafe image (Δsafety = 0.25), where overhead cables are removed and grass and trees added, the interpretation reads: *"The overhead cables should be removed, and adding grass and trees along the road will help increase people's sense of safety."* For a mildly unsafe image (Δsafety = 0.04), with only garbage bags removed and greenery and road slightly added, the model notes: *"This street is already in good condition, but by incorporating more greenery and eliminating garbage bags, it could be improved even further."*

The explanations calibrate their urgency to the magnitude of the perception gap: the largest-gap case produces the most prescriptive intervention, while the smallest-gap case acknowledges the scene's existing quality. This behavior emerges from the grounded prompting strategy — the LLM responds to the actual counterfactual difference vector, not to a general scene description — and provides evidence that the pipeline produces contextually appropriate recommendations rather than generic urban design advice.

5.5 Limitations

Several limitations of this work should be acknowledged. First, the geographic scope is restricted to Rio de Janeiro, and findings regarding which disorder elements matter most may not generalize directly to cities with different built environments or

sociocultural norms. Future work should validate the approach across multiple cities to assess cross-cultural robustness. Second, manual pixel-level annotation of 13 disorder elements is time-consuming and subject to annotator variability. Scaling this approach to larger image collections would require semi-supervised labeling strategies or automated disorder detection models.

Third, graffiti is annotated here as a single disorder category, without distinguishing tags and vandalism from muralism and commissioned street art. This is a simplification: not all graffiti signals neglect, and artistic or sanctioned murals may instead be perceived as markers of care, identity, or vitality and read as *safer* rather than more disordered. By collapsing these cases into one class, the present work may under- or over-estimate the perceptual weight of graffiti depending on scene context. This conflation, and its implications for the disorder vocabulary, is examined further in Chapter 8.

Fourth, counterfactuals in UrbanCFX are generated in feature space rather than image space: they specify changes to visual element ratios or presence flags but do not produce edited images. The proposed interventions therefore cannot be visually validated for realism or perceptual coherence by a human observer. This limitation is directly addressed by UrbanCounterViz, introduced in Chapter 6, which generates photorealistic edited images of the proposed changes.

Fifth, while the LLM interpretation layer produces fluent and grounded explanations, language models can occasionally overstate causality or introduce inferences not strictly warranted by the difference vector. The structured prompting strategy mitigates this risk, but the resulting interpretations should be understood as natural language renderings of model-derived counterfactuals rather than verified causal accounts.

Finally, the safety perception scores underlying all classification and counterfactual experiments are derived from crowdsourced pairwise judgments, which reflect aggregate public opinion rather than the perspectives of local residents or perceived safety measures. This subjectivity introduces noise and may be partially addressed in future work by integrating local surveys, crime data, or demographic context.

5.6 Final Considerations

This chapter presented UrbanCFX, a pipeline that advances urban safety perception analysis in two complementary directions. Empirically, the UrbanPD4k dataset and its 13 disorder annotations show that physical disorder elements — overhead cables, graffiti, structural damage — carry stronger predictive and counterfactual signal for perceived unsafety than general segmentation categories, and that pixel-ratio representations provide richer analytical grounding than binary presence indicators. Methodologically, the LLM interpretation layer constitutes the primary contribution: by converting counterfactual

difference vectors into natural language, it closes the communicative gap between machine-interpretable outputs and recommendations that can support early-stage reasoning for urban planners and policymakers.

Alongside Chapter 4, this chapter establishes two complementary perspectives on interpretable urban safety analysis. Chapter 4 grounds explanation in *spatial model attention* — which image regions influence the prediction — through Grad-CAM and segmentation IoU. The present chapter grounds explanation in *counterfactual intervention* — which element changes would shift a prediction — through counterfactuals, and externalizes those interventions as human-readable language. Both perspectives, however, operate on static feature representations and support neither interactive exploration of scenes nor photorealistic visualization of the proposed changes.

These remaining limitations directly motivate the central system contribution of this thesis. UrbanCounterViz, introduced in Chapter 6, builds on the visual element analysis and disorder vocabulary established in these two chapters, incorporates photo-realistic VLM-driven scene editing, and embeds the full pipeline in an interactive visual analytics environment. Where UrbanCFX produces a text sentence describing a suggested intervention, UrbanCounterViz renders a modified SVI and allows the analyst to trace, compare, and interrogate the transformation step by step. In doing so, UrbanCounterViz also adopts and extends the semantic vocabulary developed in UrbanPD4k: the disorder element categories identified here — overhead cables, graffiti, damaged walls, broken sidewalks — inform the semantic difference analysis and editing action generation at the core of UrbanCounterViz's path-guided counterfactual pipeline, introduced in Chapter 6.

6 UrbanCounterViz: A Visual Analytics System for Grounded Urban Safety Exploration

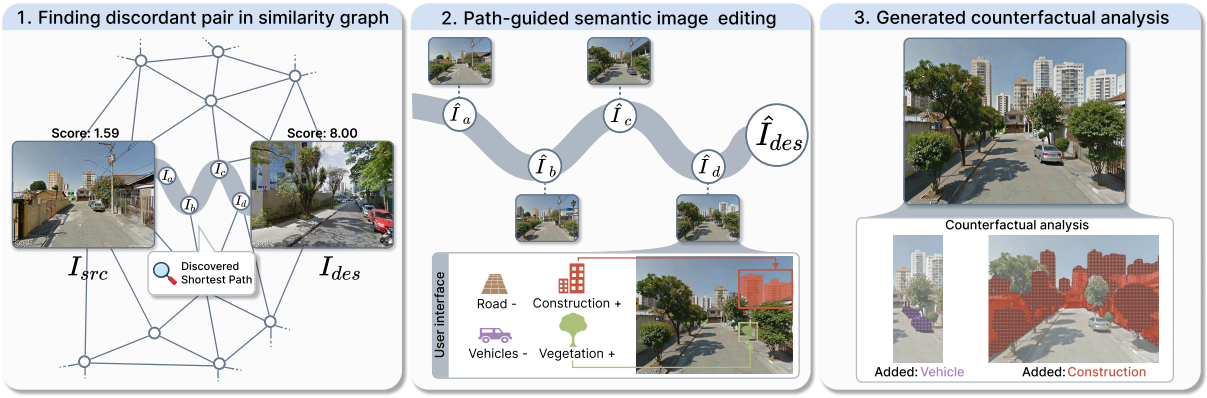

Figure 19 – Overview of UrbanCounterViz, summarizing the main stages of the visual analytics workflow for safety-perception counterfactual analysis. (1) A source image and a destination image with discordant safety-perception scores are selected from the similarity graph, and the shortest path between them is computed. (2) Semantic differences along this path guide iterative image generation toward a safety-perception increase or decrease goal. (3) The generated counterfactuals are then analyzed through image composition, urban indicators, and interactive visualizations of the edits introduced at each step. Source: the author.

6.1 Introduction

The preceding chapters established two complementary but incomplete perspectives on interpretable urban safety analysis. Chapter 4 demonstrated that specific visual elements — trees, sidewalks, vegetation, walls, bare ground, and trash receptacles — consistently shape model attention during safety prediction, but its explanations are spatial and static: they reveal *where* a model looks, not *what an analyst should change*. Chapter 5 took a step toward actionability by generating counterfactual difference vectors and converting them into natural language recommendations, but its interventions exist only in feature space — they specify *what* should change without showing *how* those changes would look in a real urban scene. Both approaches also share a deeper limitation: they are non-interactive pipelines that produce a single output for a given input, with no mechanism for the analyst to explore alternative transformations, inspect intermediate steps, or relate visual changes to perceptual outcomes in a unified environment.

This chapter introduces UrbanCounterViz, a visual analytics system designed to close these gaps simultaneously. Rather than replacing the feature-level reasoning established in Chapters 4 and 5, UrbanCounterViz extends it into the image domain: semantic differences between real urban scenes are translated into photorealistic edits via a two-stage VLM pipeline, and the entire transformation process is made inspectable through an interactive visual analytics interface. Figure 19 overviews the proposal.

Visualization and visual analytics are well suited to this challenge because they enable the integrated exploration of image content, semantic structure, model behavior, and domain knowledge. Concurrently, counterfactual and generative image-editing approaches offer a promising direction by illustrating how an image might be transformed to alter a model outcome. However, these methods frequently optimize edits based on model responses or open-ended prompt-based generation, rather than grounding them in transformations observed in real urban scenes, and consequently may produce visually striking results that are difficult to interpret as realistic urban interventions. UrbanCounterViz is designed to occupy exactly this gap: rather than detecting elements in isolation or generating edits from an unconstrained prompt, it constructs *grounded counterfactual urban edits* from *semantically meaningful differences* identified between real images with varying perception scores, so that every proposed intervention is supported by evidence already present in the data rather than by a model's unconstrained imagination.

Contributions. This chapter contributes the third and final empirical contribution of this thesis (Contribution 4, as enumerated in Chapter 1): (i) a path-guided counterfactual editing pipeline that models urban scenes as nodes in a similarity graph, identifies paths between visually similar scenes with different perception scores, and derives semantically driven edits from the object-level differences observed along these paths; (ii) UrbanCounterViz itself, a visual analytics system that orchestrates counterfactual generation, visual highlighting of edited regions, semantic summaries, and coordinated views to support human-in-the-loop analysis of how scene changes relate to perceived safety; and (iii) a reusable interpretable visual analysis workflow, evaluated in Chapter 7, that shows how the resulting counterfactuals support reasoning about the semantic factors associated with perceived safety and the identification of plausible strategies for scene improvement.

6.2 Related Works

This chapter's contribution sits at the intersection of two research areas not yet covered in this thesis: visual analytics for urban image analysis, and generative methods for urban scene editing. Chapters 4 and 5 already positioned this work against computational models for urban perception — appearance-based and segmentation-based classifiers, and counterfactual explanation methods, respectively — so this section focuses on the two

areas specific to UrbanCounterViz itself.

6.2.1 Visual Analytics for Urban Image Analysis and Perception

Visualization and visual analytics have played an important role in helping planners and analysts interpret large-scale urban imagery. *StreetVizor* (Shen et al., 2017) introduced an interactive visual analytics system for exploring human-scale urban forms from SVI, supporting comparison across areas of interest and multiple spatial scales. *Urban Mosaic* (Miranda et al., 2020) extended this direction by enabling retrieval, clustering, and comparison over large, spatially and temporally dense street-view collections, helping practitioners analyze urban change and compare neighborhoods. More broadly, VIS4ML (Sacha et al., 2019) characterizes the design space for visual analytics-assisted machine learning, and systems such as *ExplainExplore* (Collaris and van Wijk, 2020) illustrate how comparison and coordinated views support model understanding; recent surveys emphasize the importance of integrated human-in-the-loop workflows for explainable deep learning (Choo and Liu, 2018; La Rosa et al., 2023).

Despite these contributions, prior visual analytics systems for urban imagery have primarily focused on exploration, retrieval, and descriptive auditing of street-view data — supporting analysts in understanding *what* is present in a scene or *how* elements are distributed across a city, but not reasoning about *what could be changed* to produce a different perceptual outcome. Visual analytics research on explainable AI, in turn, has largely addressed tabular or generic prediction settings, without tackling the specific challenges of semantic scene editing in urban contexts. UrbanCounterViz bridges these two lines of work directly, adopting the interactive, multi-scale exploration paradigm established by urban visual analytics systems while extending it with a counterfactual analysis layer that lets analysts inspect grounded semantic edits and trace modifications back to specific image regions.

6.2.2 Generative Methods for Urban Scene Editing

Generative methods — including GANs, VAEs, diffusion models, and inpainting approaches — have been applied to urban scene editing for tasks such as visual beautification, preference-based modification, and simulation of urban renewal interventions. *FaceLift* (Joglekar et al., 2020) demonstrated that GANs could generate facade improvements that increase perceived attractiveness. UCA (Law et al., 2023) used generative models to explain changes in urban attractiveness. More recent diffusion-based work includes UPD-Explainer (Hu et al., 2023), which generates counterfactual images highlighting signs of physical disorder; VPI-Toolkit (Tan et al., 2025) and URSimulator (Hu et al., 2025), which simulate visual perception interventions. VLMs have separately become central to urban computing more broadly, used to compare machine and human judgments,

infer urban indicators from imagery, and learn multimodal representations for downstream urban tasks (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).

Despite this progress, generative approaches to urban scene editing share persistent limitations. They often provide limited control over the specific changes introduced — edits may include unrealistic architectural transformations, illumination inconsistencies, or viewpoint shifts that compromise scene fidelity — and, more critically, they offer minimal support for understanding *why* a generated change is meaningful or *how* it relates to the intended perceptual objective, since edits are optimized to change a model's output or satisfy a style criterion rather than being grounded in transformations observed in real urban environments.

Differentiation. UrbanCounterViz addresses the limitations of both lines of work simultaneously. Unlike prior urban visual analytics systems, it does not stop at exploration and descriptive auditing: it adds a counterfactual layer that lets analysts inspect grounded semantic edits and relate them to perceived safety. Unlike prior generative editing methods, it does not synthesize hypothetical modifications or optimize edits to shift a model's output: every intervention is derived from semantic differences actually observed between real, visually similar urban scenes, and the resulting edit — along with the similarity graph, the semantic difference vector, and the editing actions that produced it — is exposed to the analyst through an interactive, coordinated interface rather than delivered as an opaque final image. This combination of grounding, generation, and interactive inspection within a single environment is what distinguishes UrbanCounterViz from every system reviewed above.

6.3 System Overview

Understanding how variations in visual elements influence human safety perceptions requires more than a predictive model — it requires a framework for interactively exploring those relationships, connecting them to realistic interventions, and evaluating their plausibility. Building on the analytical pipeline established in prior chapters and on established visual analytics principles such as "*overview first, zoom and filter, then details-on-demand*" (Shneiderman, 2002), we derive a set of design requirements and analytical tasks that collectively frame the development of UrbanCounterViz.

These requirements were not defined in isolation. They were informed by the empirical analyses reported in Chapters 4 and 5, which established the visual element vocabulary, the disorder annotation framework, and the counterfactual reasoning strategy that UrbanCounterViz extends into the image domain (Moreno-Vera et al., 2024, 2025).

However, those studies also exposed a consistent limitation: reasoning exclusively through aggregated or tabular representations makes it difficult to understand how such changes would manifest visually in a real scene. This gap — between the feature-level description of an intervention and its visual appearance — motivated the transition to image-based analysis and the design of UrbanCounterViz as a system that connects semantic differences, plausible scene edits, and perceptual outcomes in a visually interpretable and interactive environment.

6.3.1 Design Requirements

Based on the research gaps identified in Chapter 3 and on the lessons learned from the empirical contributions in Chapters 4 and 5, we established four core requirements (R) to guide the design of UrbanCounterViz.

R1 — Contextualize Scene Similarity and Transitions. Users should be able to select geo-located SVI, compare visually similar urban scenes, and inspect transition sequences connecting them. The system should reveal how perceptual attributes and semantic categories evolve across similar scenes and along the path between two selected images. This requirement responds directly to the absence of interactive exploration in prior urban perception systems (Shen et al., 2017; Miranda et al., 2020), which support descriptive auditing of scene collections but not reasoning about how scenes transition into one another.

R2 — Steer Semantic Edits. Users should be able to inspect semantic differences between selected images and use these differences to guide counterfactual generation. The system should provide planning-oriented semantic edits and support the generation of guided counterfactual images. This extends the counterfactual reasoning introduced in Chapter 5 from feature space into the image domain, where interventions can be visually evaluated.

R3 — Inspect and Understand Visual Modifications. When a counterfactual image is generated, users should be able to identify what changed, where it changed, and how those changes relate to the intended perceptual shift. The system should support detailed visual comparison while clearly distinguishing added, removed, and preserved semantic categories. This requirement addresses the opacity of prior generative methods (Joglekar et al., 2020; Hu et al., 2023, 2025), which produce edited images without revealing the transformation rationale.

R4 — Evaluate Multi-dimensional Impact. Users should be able to assess the effects of edits across multiple dimensions simultaneously. The system should provide visual summaries that compare perceptual attributes (e.g., safety perception scores), urban indicators (e.g., openness, greenery, walkability), and semantic category distributions

before and after editing. This requirement connects to the broader need for tools that support transparent, evidence-grounded urban design recommendations, as identified in the final considerations of Chapter 3.

6.3.2 Analytical Tasks

Building on these requirements, we define a concise set of analytical tasks that UrbanCounterViz is designed to support. These tasks follow an overview-to-detail pattern (Shneiderman, 2002): analysts initially obtain a broad spatial overview of the image collection, then examine similarities and perceptual scores across selected scenes, and finally inspect the details of the editing process and its outcomes.

T1: Explore scene similarity and transitions (R1).

- **T1.1** Identify and compare visually similar geo-located scenes based on perceptual attributes and semantic composition.
- **T1.2** Analyze transition sequences between two selected images to understand how semantic categories and perceptual scores change along the path.

T2: Analyze and steer semantic edits through counterfactual generation (R2).

- **T2.1** Identify and quantify semantic differences between a pair of selected images and their generated counterparts.
- **T2.2** Inspect intermediate transition steps in the generation process to understand how semantic categories and perceptual attributes progressively change and where key differences emerge.

T3: Compare, interpret, and relate multi-dimensional impacts (R3, R4).

- **T3.1** Identify and characterize visual modifications — added, removed, and preserved semantic categories — and relate them to perceptual changes.
- **T3.2** Relate urban indicators, semantic categories, perceptual scores, and similarity patterns across original and generated counterfactual images.

These tasks directly informed the design of UrbanCounterViz's six coordinated views, each of which addresses specific requirements while remaining integrated with the others to enable seamless exploration, comparison, and interaction.

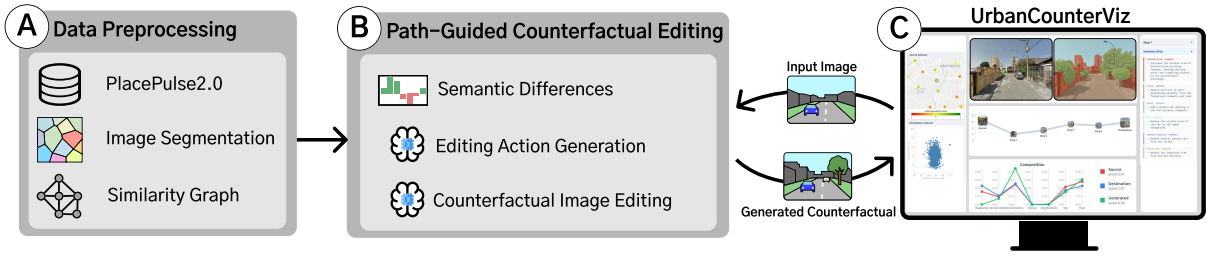

Figure 20 – Overview of the URBANCOUNTERVIZ workflow. (A) **Data preprocessing**: we transform Place Pulse 2.0 pairwise comparisons into structured representations, including perception scores, semantic categories, deep visual features, and image similarities. (B) **Path-guided counterfactual editing**: semantic differences between image pairs are used to derive editing actions and generate grounded counterfactual images. (C) **UrbanCounterViz**: an interactive visual analytics system that supports exploration of urban scenes, inspection of semantic changes, comparison of generated edits, and evaluation of how these edits relate to perceived safety. Source: the author.

6.3.3 Workflow

Figure 20 provides a high-level overview of the URBANCOUNTERVIZ workflow, which consists of three stages executed in sequence.

In the **dataset preprocessing stage** (Figure 20.A), raw SVI and pairwise human judgments are transformed into structured representations: perception scores, deep visual feature embeddings, semantic category proportions, urban indicators, and a pairwise image similarity graph. This stage runs offline and produces the data structures that support all subsequent interaction.

In the **path-guided counterfactual editing stage** (Figure 20.B), the system computes semantic differences along graph paths between a user-selected source image and a destination image with a higher safety perception score. These differences are used to derive explicit editing actions, which are then passed to a two-stage VLM pipeline to generate photorealistic modified images. Crucially, this stage decomposes a potentially large perceptual gap into smaller, interpretable steps, each grounded in differences observed between real adjacent scenes in the graph.

Finally, the **visual analytics interface** (URBANCOUNTERVIZ, Figure 19.C) integrates these computational components into an interactive environment where users can guide image generation, inspect semantic and visual changes at each step, and evaluate whether the resulting counterfactuals support their analytical goals. The interface is the primary contribution of the system: it makes the pipeline transparent, steerable, and interpretable rather than a black-box transformation from input to output.

6.4 Data Preprocessing

The preprocessing stage converts raw pairwise human judgments and SVI into the structured representations required for interactive exploration and counterfactual generation. It comprises three steps: (i) estimating image-level perceptual scores from pairwise comparisons, (ii) extracting semantic features and computing urban indicators from images, and (iii) constructing a similarity graph over the image collection.

6.4.1 Perception Score Estimation

Pairwise comparison outcomes are converted into image-level perceptual scores using the Strength of Schedule (SOS) ranking algorithm, formally introduced in Section 2.4.2. This approach produces Q-scores, scaled from 0 (very unsafe) to 10 (very safe), which serve as the perceptual foundation for all subsequent scene selection, path computation, and counterfactual evaluation within UrbanCounterViz. Although Q-scores are computed globally across all images, the subsequent analysis and graph construction are performed independently for each city.

6.4.2 Semantic Feature Extraction and Urban Indicators

To characterize the visual content of each image, semantic category proportions are extracted using OneFormer pre-trained on ADE20K (Jain et al., 2023) — the same segmentation model established in Chapters 4 and 5. For each image I, the proportion of pixels $p_k(I)$ assigned to each of the 150 ADE20K semantic classes is computed. These proportions provide a concise, comparable summary of scene composition and are subsequently used to identify semantic differences between images and to derive candidate editing actions.

Because 150 individual categories are too fine-grained for interactive analysis and inspection, they are aggregated into nine higher-level groups: **Sky**, **Human**, **Construction**, **Floor**, **Vegetation**, **City elements**, **Terrain Vehicles**, and **Other**. This grouping preserves the semantic distinctions most relevant to safety perception — as established in Chapter 5 — while presenting them at a level of abstraction that analysts can readily interpret and act on. The **Other** category is excluded from interactive analysis due to its minimal and heterogeneous representation in the dataset.

In addition to semantic proportions, six higher-level urban perception indicators are computed for each image: **greenness**, **walkability**, **driving safety**, **imageability**, **enclosure**, and **openness** (Ma et al., 2021; Qiu et al., 2023). These indicators provide a complementary layer of analysis that goes beyond object-level composition to capture environmental qualities commonly associated with human perceptions of urban spaces.

They are displayed alongside semantic proportions in the interface to support multi-dimensional evaluation of generated edits (R4).

6.4.3 Discordant Pairs in Urban Perception

A recurring challenge in urban perception research is understanding *why* two visually similar scenes receive substantially different human perception scores. Standard prediction pipelines produce a score per image but offer no direct mechanism for identifying which visual differences are responsible for divergent human judgments. To address this, we introduce the concept of **discordant pairs** as a formal contribution of this thesis. Although the intuition behind comparing similar scenes with different scores motivated the visual element analysis in Chapter 4 and the counterfactual approach in Chapter 5, this chapter formalizes the concept and operationalizes it as a structural component of the UrbanCounterViz pipeline.

Formally, let $\mathcal{I} = \{I_1, I_2, \ldots, I_n\}$ be a set of SVI, each associated with a perception score $q_i \in [0, 10]$ estimated from pairwise comparisons (e.g., via Elo, TrueSkill, or Strength of Schedule, as described in Section 2.4.2). Let $\phi : \mathcal{I} \to \mathbb{R}^d$ denote a feature embedding function that maps each image to a d-dimensional representation. A **discordant pair** (I_i, I_j) is defined as a pair of images satisfying two conditions simultaneously:

- **Visual similarity**: the images are close in the embedding space, i.e., $sim(\phi(I_i), \phi(I_j)) \geq \tau$ for some $\tau > 0$;
- **Perceptual divergence**: their perception scores differ substantially, i.e., $|q_i - q_j| \geq \sigma$ for some $\sigma > 0$.

The concrete values used in this work are reported in Section 6.7. Intuitively, discordant pairs are scenes that look alike yet are judged very differently by human observers. Because the two images share most of their visual content, any difference in perception can be attributed to a relatively small set of distinguishing elements rather than to global scene variation. This property makes discordant pairs particularly informative for counterfactual analysis: by systematically examining the semantic and object-level differences within such pairs, it is possible to identify which visual elements — such as the presence of vegetation, lighting conditions, building facades, or signs of physical disorder — are most strongly associated with shifts in perceived safety.

Whereas counterfactual explanations ask *what minimal change to an input would alter a model's prediction*, discordant pairs ground this question in real observed data: instead of synthesizing hypothetical modifications, the analysis is driven by differences that actually exist between real urban scenes. The resulting interventions are therefore more plausible and directly interpretable as realistic urban conditions. This grounded

counterfactual strategy forms one of the core methodological contributions of the present work and is discussed in greater detail in Section 6.5.

6.4.4 Similarity Graph Construction

To capture visual relationships between images at scale, an undirected similarity graph $G = (\mathcal{V}, \mathcal{E})$ is constructed, where each node $v \in \mathcal{V}$ corresponds to a SVI and each edge $(v_i, v_j) \in \mathcal{E}$ connects images that are visually similar. For each image, a deep feature vector is extracted using a ResNet101 model pre-trained on ImageNet, without fine-tuning on perceptual quality scores, ensuring that feature representations capture general visual patterns independently of the scores. This keeps graph topology and score differences as two decoupled sources of information, which is essential for meaningfully identifying discordant pairs. Pairwise cosine similarities are then computed across the collection, and two images are connected by an edge whenever their cosine similarity exceeds a threshold τ, retaining only strongly similar pairs.

This graph formalizes the concept of discordant pairs introduced in Section 6.4.3: pairs of nodes that are close in the embedding space (connected by an edge or reachable via a short path) but differ substantially in their safety perception scores. Within UrbanCounterViz, the graph serves two complementary purposes. First, it provides the structural basis for the *Destination Selector* view, which presents candidate target images reachable from a selected source image. Second, and more centrally, it defines the paths along which counterfactual edits are computed: the shortest path between a source node and a destination node with a higher safety perception score decomposes a potentially large perceptual gap into a sequence of smaller, visually grounded transitions, each informed by differences observed between real adjacent scenes.

By integrating perception scores, semantic summaries, and graph-based visual similarity, the preprocessing stage produces the representations that enable all subsequent interactive exploration and grounded counterfactual generation in UrbanCounterViz.

6.5 Path-Guided Counterfactual Editing Pipeline

The path-guided editing pipeline is the computational core of UrbanCounterViz. Given a source image I_{src} selected and a destination image I_{des}, the objective is to generate a grounded counterfactual image that transitions the source toward the semantic composition of the destination while preserving the structural coherence and photorealistic appearance of the original scene. This is achieved through three sequential components: path guidance, semantic difference modeling, and two-stage VLM-driven image editing.

6.5.1 Path Guidance

For a selected pair (I_{src}, I_{des}), the shortest path is computed between their corresponding nodes in the similarity graph defined in Section 6.4.4. This path produces a sequence of intermediate images that are each visually similar to their neighbors and that collectively bridge the source and destination:

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

The path is central to the interpretability of the approach. Rather than requiring the editing model to transition directly from the source to a potentially distant target — which would risk producing unrealistic or arbitrary modifications — the transformation is decomposed into smaller, contextually grounded steps. Each step is guided by differences actually observed between nearby real urban scenes, not by unconstrained model optimization or free-form prompting. This decomposition connects directly to the key limitation identified in Chapter 5: UrbanCFX produced counterfactuals in feature space that could not be visually validated. Here, each step along the path is itself grounded in a real image, ensuring that every generated edit is anchored in an existing urban condition.

Generation proceeds sequentially along the path: at each step, the current image is edited to produce an updated image that serves as the source for the next step. Semantic edits therefore accumulate incrementally, and the analyst can inspect the transformation at any intermediate point rather than only examining the final output.

6.5.2 Semantic Difference Modeling

At each step of the path, the semantic composition of the current source image is compared with that of the next target image using the category proportions defined in Section 6.4.2. Let $\hat{I}_{\mathcal{H}}$ denote the current source image and X the next target image along the path. For each semantic category k, the raw signed difference is computed as:

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

and normalized as:

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

yielding $\delta_k \in [-1, +1]$, where positive values indicate categories that should increase, negative values indicate categories that should decrease, and values near zero indicate little or no change. The full semantic difference vector is then:

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

Not every semantic difference is equally meaningful for editing. Only categories with $|\delta_k|$ above a minimum threshold are retained and treated as actionable categories for the

current step. This filtering serves two purposes: it reduces noise from minor compositional fluctuations between similar scenes, and it focuses the editing actions — and the analyst's attention — on the changes most likely to produce a meaningful perceptual shift. The filtered semantic difference vector $\boldsymbol{\delta}$ forms the direct input to the image editing stage and is also displayed in the interface to support transparency and human oversight (R2, R3).

6.5.3 Two-Stage VLM-Driven Image Editing

Counterfactual image generation is implemented through a two-stage vision-language workflow that separates *what should change* from *how the image is edited*. This separation is an intentional design decision that enhances interpretability and transparency: by producing an explicit intermediate representation of the intended transformation, the pipeline allows analysts to inspect, understand, and if needed challenge the edit rationale before examining the generated image.

Stage 1 — Edit Action Generation. In the first stage, a set of candidate editing actions is derived from the semantic difference vector, the current source image, and the safety perception scores of the current and target states. A VLM processes the source image alongside structured semantic guidance and returns a set of edit actions expressed as explicit visual instructions — for example, *increase tree coverage along the sidewalk*, *reduce the prominence of overhead cables*, or *introduce maintained pavement markings*. These actions are bounded by the semantic difference vector and formally defined as:

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

The resulting action set is displayed in the *Generation View* of the interface, where analysts can review it before inspecting the generated image. Prompt 2 details all the instructions, context, and reasoning flow.

```
[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 used to generate a list of actions

```
##  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 – (continued)

Stage 2 — Photorealistic Image Editing. In the second stage, the edit actions $\mathcal{E}_{\mathcal{H}}^{X}$ and the current source image are passed to a second VLM, which produces the edited image as:

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

applying the requested modifications while preserving the remaining scene content, spatial structure, viewpoint, and overall visual coherence of the original scene. Prompt 3 details constraints and rules to edit images, aiming to keep original visual characteristics.

```
[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 used to edit an input image following a list of actions

After generation, the edited image is re-segmented with OneFormer, and its semantic category proportions, urban indicator values, and estimated safety score are recomputed. These derived outputs allow the interface to immediately compare the generated image against the source and destination along multiple dimensions (R4), enabling analysts to assess whether the edit moved the scene in the intended perceptual direction and whether the changes are visually coherent and semantically consistent with the guiding differences.

This two-stage design enhances interpretability by making the editing process transparent and auditable. Unlike approaches that directly generate a modified image,

UrbanCounterViz explicitly separates the identification of what should change from the act of changing it, ensuring that every generated image is accompanied by the semantic differences that motivated it, the editing actions that guided it, and the recomputed perceptual and semantic summaries that evaluate it.

6.6 The UrbanCounterViz Interface

UrbanCounterViz is an interactive visual analytics system that integrates the computational pipeline described above into a unified exploratory environment. The interface comprises six coordinated views — *Source Selector*, *Destination Selector*, *Path View*, *Image View*, *Generation View*, and *Image Stats* — that collectively support the analytical tasks defined in Section 6.3.2. Figure 21 presents the full interface; Table 11 summarizes the mapping between views and tasks.

Table 11 – UrbanCounterViz views and the analytical tasks they support.

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

6.6.1 Source Selector (A)

The *Source Selector* is the entry point to the analytical workflow. It presents a map-based overview of the image collection (Figure 21A), where each point represents a geo-located SVI colored by its perceptual Q-score. This spatial encoding allows analysts to immediately identify areas of low, moderate, or high perceived safety and to contextualize their analysis within the city's geography.

Selecting a point designates the source image I_{src} to be transformed and initializes the downstream workflow: the *Destination Selector* is updated to show candidate targets reachable from I_{src} within the similarity graph, and all other views reset to reflect the new selection. This view primarily supports T1 by providing the spatial and perceptual foundation that orients the analyst before any counterfactual reasoning begins.

6.6.2 Destination Selector (B)

The *Destination Selector* presents a scatterplot of candidate images reachable from I_{src} within the similarity graph (Figure 21B). The horizontal axis encodes visual

Figure 21 – URBANCOUNTERVIZ interface. (A) **Source Selector**: geo-located images colored by perceptual Q-score; selecting a source initializes the workflow. (B) **Destination Selector**: scatterplot of candidate images showing visual similarity to the source (x) versus Q-score (y), supporting destination selection. (C) **Path View**: shortest path in the similarity graph from source to destination, exposing intermediate scenes and their semantic differences. (D) **Image View**: side-by-side comparison of the source (D.1) and generated counterfactual (D.2), augmented with overlays highlighting added and removed regions. (E) **Generation View**: stepwise editing actions organized by semantic category, enabling inspection of the generation process. (F) **Image Stats**: coordinated summaries comparing semantic composition (F.1) and urban indicators (F.2) across the source, destination, and generated images. Source: the author.

similarity to the source image and the vertical axis encodes the perceptual Q-score, creating a two-dimensional space in which the analyst can simultaneously reason about scene relatedness and perceptual distance. The upper-right region of this scatterplot is of particular analytical interest: images located there are both visually similar to the source and associated with substantially higher perceived safety — precisely the conditions that define a strong discordant pair and a high-quality candidate for counterfactual editing.

By making the similarity–safety perception trade-off visually explicit, this view supports T1 and gives the analyst direct control over the degree of visual deviation the resulting counterfactual will involve. Selecting a point in the scatterplot designates I_{des} and triggers path computation and interface updates across all downstream views.

6.6.3 Path View (C)

The *Path View* visualizes the shortest path in the similarity graph between I_{src} and I_{des} (Figure 21C), as computed in Section 6.5.1. It exposes the intermediate images that connect source and destination through visually related scenes, rendering the transformation process as a sequence of incremental transitions rather than a single abrupt change.

Nodes are positioned along the horizontal axis by pairwise dissimilarity and along the vertical axis by perceptual Q-score, enabling inspection of whether the path follows a smooth perceptual trajectory or involves non-monotonic score changes. Semantic differences between consecutive nodes are depicted as compact bar charts that summarize the most significant category changes at each step.

This view is analytically central: it supports T1 by exposing the structural relationships between similar scenes, T2 by surfacing the stepwise semantic differences that will guide editing, and T3 by enabling analysts to anticipate how perceptual scores change across the trajectory before generation begins. Selections in this view propagate to the *Image View* and *Image Stats*, allowing focused inspection of any individual step.

6.6.4 Image View (D)

The *Image View* supports direct visual comparison between the source image and the generated counterfactual (Figure 21D). By default, the left panel displays I_{src} and the right panel displays the current generated image $\hat{I}$; intermediate steps can be loaded through selections in the *Path View* or *Generation View*. To facilitate the identification of differences between visually similar images — a cognitively demanding task when changes are subtle or localized — the system overlays semantic change masks that distinguish added regions (categories that increased) from removed regions (categories that decreased). The resulting display functions as a visual diff, allowing analysts to efficiently locate modifications and assess whether they correspond to structurally plausible changes in the scene.

This masking mechanism is a direct response to Requirement R3: it makes edits explicit and spatially localized rather than leaving the analyst to detect them through unguided visual inspection. Together with the *Generation View*, it supports T2 by confirming the spatial effects of semantic steering and T3 by enabling detailed interpretation of how generated changes are realized in the image.

6.6.5 Generation View (E)

The *Generation View* presents the counterfactual editing process as an explicit sequence of generation steps (Figure 21E). For each step, the view displays the editing actions derived in Stage 1 of the VLM pipeline, organized by semantic category. This view is central to the system's interpretability: by revealing the intermediate action representations that guide the image-editing model, UrbanCounterViz makes generation transparent and auditable rather than treating it as a black box. Analysts can review these actions alongside the corresponding generated images and assess whether the stated intentions match the observed visual outcomes — or whether they should select a different path, adjust the destination, or consider an alternative trajectory.

This explicit action representation also serves a practical purpose identified by domain experts in preliminary use: it allows analysts to predict the character of an edit before examining the generated image, reducing cognitive surprise and making the system's behavior more trustworthy. The *Generation View* primarily supports T2 (understanding how semantic differences translate into editing instructions) and T3 (auditing the relationship between intended changes and observed outcomes). Selections in this view update both the *Image View* and *Image Stats* for step-level inspection.

6.6.6 Image Stats (F)

The *Image Stats* view provides a multi-dimensional summary of the editing outcome using two coordinated panels (Figure 21F) that simultaneously compare the source, destination, and generated images.

Semantic Composition Panel (F.1). This panel compares the aggregated semantic category proportions of the three images. It enables analysts to assess whether the generated image has moved toward the destination in terms of scene composition — for instance, whether vegetation has increased or construction elements have changed — and whether the observed shifts match the semantic differences that motivated the edit.

Urban Indicators Panel (F.2). This panel compares the six higher-level urban perception indicators, along with the estimated safety score, across source, destination, and generated image. Axes are normalized per dimension to facilitate comparison across heterogeneous measures. This panel allows analysts to interpret whether visual changes in the generated image correspond to meaningful shifts in environmental qualities such as openness, greenness, walkability, and enclosure — dimensions that matter for urban design decisions beyond perception score alone.

Together, these panels implement Requirement R4 by connecting visible scene edits to changes in semantic composition, urban indicators, and estimated perceived safety. This multi-dimensional view also serves as a quality check: if the generated image produces unexpected shifts in indicators not targeted by the edit (e.g., a large decrease in openness when only vegetation was intended to increase), the analyst can flag this as a signal of generation instability and select an alternative step.

6.7 Implementation Details

UrbanCounterViz employs a client–server architecture. The client is implemented in JavaScript using D3.js (Bostock et al., 2011) for interactive graphics and coordinated interactions across all six views. The server is implemented in Python using FastAPI (Ramírez, 2019), which serves preprocessed data, graph information, image assets, and generation outputs.

Deep visual features for similarity graph construction are extracted using a ResNet101 model pre-trained on ImageNet (He et al., 2016). Semantic segmentation is performed with OneFormer pre-trained on ADE20K (Jain et al., 2023), accessed through the HuggingFace ecosystem (Wolf et al., 2019). Safety perception score estimation after generation is conducted using a ConvNext model trained on Place Pulse dataset. Image editing uses Gemini `2.5-flash` and `2.5-flash-image` (Google, 2025) for Stage 1 (action generation) and Stage 2 (photorealistic editing), respectively. All preprocessing steps — including feature extraction, segmentation, score computation, and graph construction — are performed offline to ensure responsive interactions during live analysis. Only image generation is triggered on-demand during a session.

The thresholds governing the pipeline — visual similarity (τ), perceptual divergence (σ), and semantic-difference filtering — control the trade-off between grounding and diversity. Stricter values produce pairs that are more visually similar and easier to interpret, but reduce the number of available transformations; looser values expand the search space at the cost of less plausible edits. In this work, these parameters were selected empirically to favor interpretability and visual coherence over exhaustive coverage: e.g., $\tau = 0.85$ for edge creation, $\sigma = 4.0$ Q-score points for perceptual divergence, and $|\delta_k| > 0.10$ for semantic-difference filtering.

6.8 Final Considerations

This chapter introduced UrbanCounterViz, a visual analytics system that addresses the principal limitations identified across the prior chapters of this thesis. Where Chapter 4 produced static spatial attributions and Chapter 5 produced text-only feature-space recommendations, UrbanCounterViz generates photorealistic, grounded, and interactively explorable urban scene transformations. By grounding every edit in differences observed between real adjacent scenes in a similarity graph, the system avoids the principal weakness of prior generative approaches — the production of visually striking but analytically uninterpretable modifications. By making the full transformation process — from path selection to semantic difference computation to edit action generation to photorealistic output — transparent and inspectable within a coordinated visual analytics interface, it supports the kind of human-in-the-loop reasoning that urban planning decisions require.

Two properties of the design deserve emphasis as they distinguish UrbanCounterViz from all prior related work reviewed earlier in this chapter. First, grounding: interventions are derived from patterns that exist in real data, not from model optimization or unconstrained prompting, making them more credible as evidence for design decisions. Second, transparency: every stage of the pipeline is exposed to the analyst through the interface, from the similarity graph and candidate destinations to the semantic difference

vectors, editing actions, generated images, and multi-dimensional outcome summaries. Together, these properties support not only the generation of counterfactual urban scenes, but the interpretable, evidence-grounded reasoning about those scenes that constitutes the central contribution of this thesis.

Chapter 7 presents the evaluation of UrbanCounterViz through three usage scenarios, an expert feedback study, and user feedback study, examining how the system supports analysis across different urban conditions, trajectory lengths, and directions of transformation.

7 UrbanCounterViz Evaluation

The preceding chapter described how UrbanCounterViz turns the reasoning built up over the last three chapters into something an analyst can actually work with: a scene, a plausible edit, and a way to see it. What remained untested was whether it actually works as intended for the people meant to use it. This chapter asks that question directly, through the eyes of the people who would use such a system — planners, analysts, and everyday observers of a street.

The evaluation is organized into three complementary parts, each addressing a different dimension of the central evaluative question: does UrbanCounterViz support transparent, human-in-the-loop reasoning about plausible urban transformations in ways that neither prior visual analytics systems nor prior generative methods have provided? The first part presents three usage scenarios conducted with SVI from São Paulo, each targeting a different analytical goal: transforming a low-safety scene into a visually similar but higher-safety one, exploring the behavior of longer multi-step trajectories, and examining counterfactual generation in the reverse direction toward lower perceived safety. These scenarios are not intended as formal user experiments but as structured demonstrations of the system's analytical range — what it reveals, where it excels, and where its outputs become less reliable. The second part reports qualitative feedback collected through semi-structured interviews with domain experts in urban planning and urban analysis, providing an external perspective on the system's usability, usefulness, and interpretability from practitioners capable of judging the realism and practical relevance of the generated edits. The third part reports a structured user study in which a broader sample of participants evaluated the system using standardized usability instruments and targeted perception questions, providing quantitative evidence that complements the qualitative expert judgment.

Together, the three parts address that question from increasingly broad angles: the scenarios demonstrate what the system can do, the expert interviews test whether domain practitioners find it credible, and the user study checks whether that credibility holds for a wider, non-specialist audience.

7.1 Usage Scenarios

The usage scenarios use SVI from São Paulo, Brazil, selected from the Place Pulse 2.0 dataset. São Paulo provides a large and visually diverse collection of urban scenes spanning a wide range of perceived safety scores, infrastructure conditions, and semantic compositions — making it well-suited for demonstrating the analytical range of the system

across different types of urban environments. Three scenarios are presented in order of increasing complexity: a focused two-step transformation, a longer monotonic trajectory, and a trajectory in the reverse perceptual direction.

7.1.1 Scenario 1 — Enhancing Urban Safety Perception Through Discordant Pairs

Figure 22 – Usage scenario showing a path-guided counterfactual transformation from a lower-safety to a safer urban scene. Starting from a low-safety source image S (score: 1.58) and a visually similar but higher-safety destination image D (score: 6.11, similarity: 0.80), UrbanCounterViz computes a path with one intermediate node (I.1) and generates two grounded edits (G.1 and G.2). The corresponding action prompts (P.1 and P.2) and semantic masks (M.C and M.V) help explain how the transformation unfolds. Source: the author.

This scenario illustrates the primary analytical goal of UrbanCounterViz: transforming a low-safety urban scene into a visually similar but higher-safety alternative, guided by differences observed in a real discordant pair.

Setup. The analyst begins in the *Source Selector* by identifying a low-scoring image in a residential area of São Paulo (Figure 22S, Q-score: 1.58). The scene shows a narrow street with limited vegetation, sparse street furniture, and few visible signs of activity — visual characteristics consistently associated with lower perceived safety in the analyses of Chapters 4 and 5. In the *Destination Selector*, the analyst identifies a visually similar image with substantially higher perceived safety (Figure 22D, Q-score: 6.11, similarity: 0.80). This pair satisfies the formal definition of a discordant pair introduced in Section 6.4.3: two scenes that are close in the visual embedding space yet differ strongly in their human-assigned safety scores.

Path and edit generation. The *Path View* reveals a short path with one intermediate node (Figure 22I.1), decomposing the transformation into two editing steps. For each step,

the pipeline derives explicit editing actions (Figure 22P.1 and P.2) from the semantic differences computed between adjacent nodes along the path. These actions include increasing tree coverage along the sidewalk, modifying visible street-side structures, and introducing cues associated with better-maintained and more active environments. The semantic category differences driving each step are displayed in the *Path View* as compact bar charts (Figure 22T.1), giving the analyst a clear preview of what will change before examining the generated images.

Results and interpretation. The final generated image (Figure 22G.2) increases the estimated safety perception score by approximately 4.5 points relative to the source, while preserving the original street layout, building structure, and viewpoint. Semantic change masks (Figure 22M.C and M.V) highlight the most relevant modifications in the *construction* and *vegetation* categories, enabling the analyst to verify that the changes are spatially localized and structurally coherent rather than globally inconsistent. The *Image Stats* view confirms that the generated image moves toward the destination in both semantic composition and urban indicators, including greenness and openness.

This scenario demonstrates a key property of the grounded counterfactual approach: rather than prompting a generative model with a free-form description of a desired scene, the edits are derived from differences actually observed between two real urban environments. The result is not only visually more coherent but also more analytically credible — the analyst can trace exactly which differences in the real pair motivated each edit, and can inspect whether the generated image realizes those differences faithfully.

7.1.2 Scenario 2 — Multi-Step Trajectory Analysis

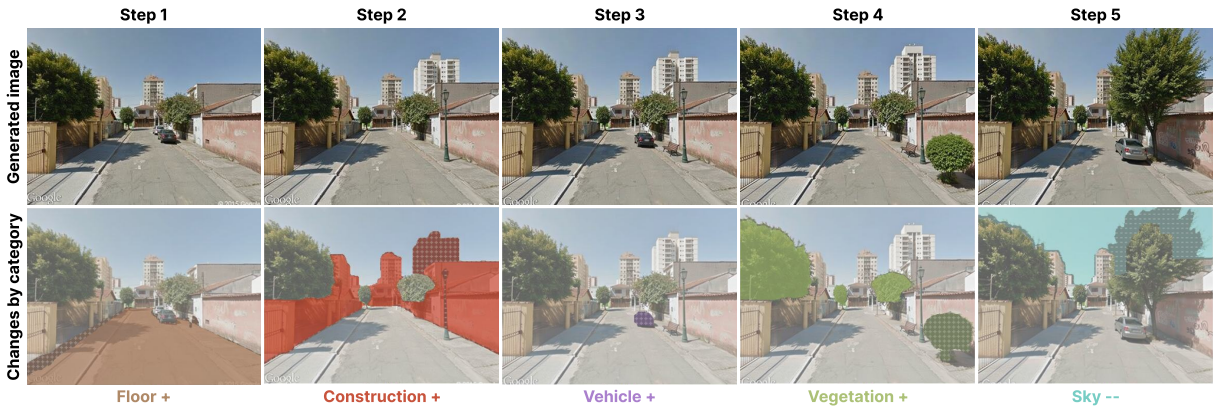

Figure 23 – Usage scenario showing a multi-step counterfactual trajectory toward higher perceived safety. Each column corresponds to one generation step, with the generated image on top and the dominant semantic change mask below. Highlighted regions indicate the class most strongly modified at each step; "+" denotes an increase in presence and "–" a strong decrease. The first step is computed relative to the source image, and subsequent steps are computed relative to the previously generated image. Source: the author.

This scenario explores the behavior of UrbanCounterViz along longer counterfactual trajectories, examining both what is gained and what is lost as the number of editing steps increases.

Setup. Starting from the same low-scoring source image as Scenario 1, the analyst selects a more perceptually distant destination (Figure 21C, Q-score: 8.0, similarity: 0.71) — placing it among the highest-scoring scenes in the São Paulo subset. A monotonic path is then identified, requiring the perceived safety score to increase at every step. The resulting trajectory passes through six nodes and involves five sequential generated images (Figure 23).

Progressive transformation. Along the trajectory, the pipeline progressively introduces elements commonly associated with higher perceived safety: additional built structure, street furniture, expanded vegetation, and changes in visible sky. The *Image Stats* view (Figure 21F) confirms that the final generated image moves substantially closer to the destination in both semantic composition and estimated safety score. Crucially, the multi-step trajectory reveals *how* these changes accumulate over time: the analyst can inspect each intermediate transformation rather than only the final result, and can identify at which step a particular category first changes meaningfully.

Diminishing returns and failure modes. The trajectory also reveals important limitations of extended generation. As shown in Figure 23, the largest gains occur during the first two or three steps. Early edits introduce visually salient and globally coherent modifications, such as expanded vegetation, added built elements, and a more structured street environment. In later steps, the estimated safety score continues to increase, but gains become smaller and more localized. Two recurrent failure modes emerge. First, the trajectory may enter an **oscillatory regime**, in which edits in later steps partially overwrite or counteract modifications introduced in earlier steps. This behavior is visible in repeated changes to the *vehicles* category in Steps 3 and 5 (Figure 23), where the pipeline alternately adds and removes vehicle elements without converging on a stable semantic state. Second, some edits become geometrically implausible as localized modifications: the insertion of large structural elements that do not integrate naturally with the surrounding street layout. These effects are made legible through the semantic change masks and *Image Stats* summaries, allowing the analyst to identify the inflection point at which generation becomes unreliable.

Comparison with Scenario 1. The contrast between the two scenarios clarifies a fundamental trade-off. Few-step editing, as in Scenario 1, yields more interpretable and visually plausible interventions — adding vegetation, improving maintenance cues, adjusting built elements — that align with realistic urban design actions. Longer monotonic trajectories expand the space of achievable perceptual scores but introduce increasing visual inconsistency and semantic instability. To test whether this observation is specific

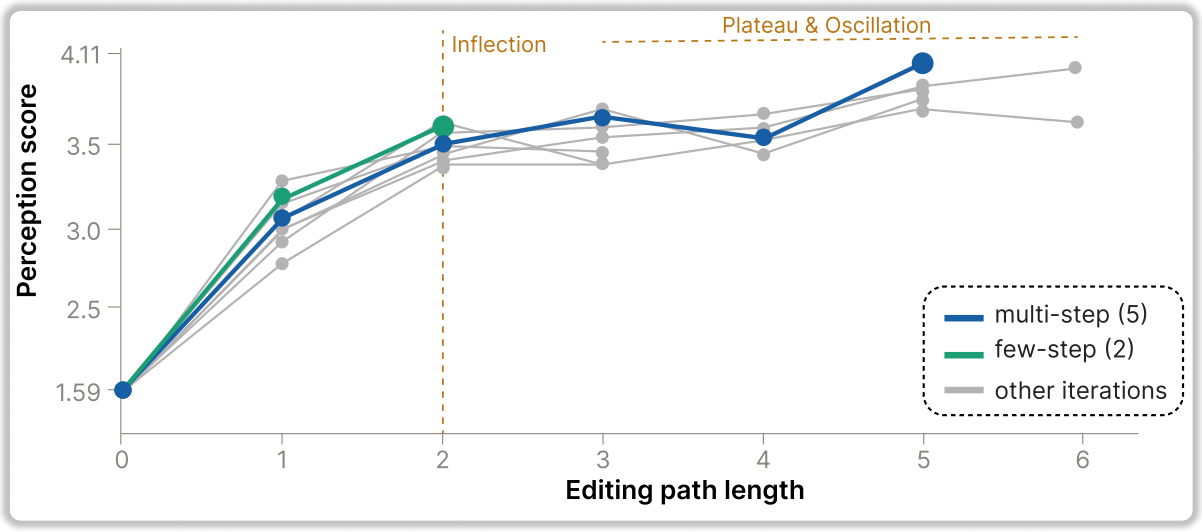

Figure 24 – Estimated perception score as a function of editing path length for the multi-step (blue) and few-step (green) trajectories, alongside additional runs from the same source image (gray). All runs exhibit consistent behavior: rapid early gains, an inflection around step 2–3, and a plateau with oscillation in later steps. Source: the author.

to a single path, additional trajectories from the same source image were computed using alternative monotonic paths toward comparably low-similarity destinations (see Figure 24). Across these runs, the pattern is consistent: rapid gains in the initial steps, an inflection point around Steps 2–3, and diminishing returns with oscillatory behavior thereafter. This suggests that the trade-off between short and long trajectories reflects a structural property of the pipeline rather than an artifact of any single path choice.

This finding highlights the practical value of the human-in-the-loop design of UrbanCounterViz: rather than generating the full trajectory automatically and presenting only the endpoint, the system allows analysts to inspect each step, identify the inflection point, and stop or redirect the trajectory when outputs begin to diverge from their analytical goals.

7.1.3 Scenario 3 — Progressive Decrease in Perceived Safety

The third scenario examines the reverse direction of the transformation process. Rather than improving a low-safety scene, the analyst uses UrbanCounterViz to progressively transform a relatively safe-looking street into one perceived as less safe. This bidirectional capability has a distinct analytical purpose: it allows researchers and planners to identify which visual cues — when introduced or removed — most strongly signal disorder, neglect, or reduced safety to human observers.

Setup. The source image is a scene with a mid-to-high safety score (Figure 25S, Q-score: 5.48), showing a maintained residential street with visible vegetation and a decorated

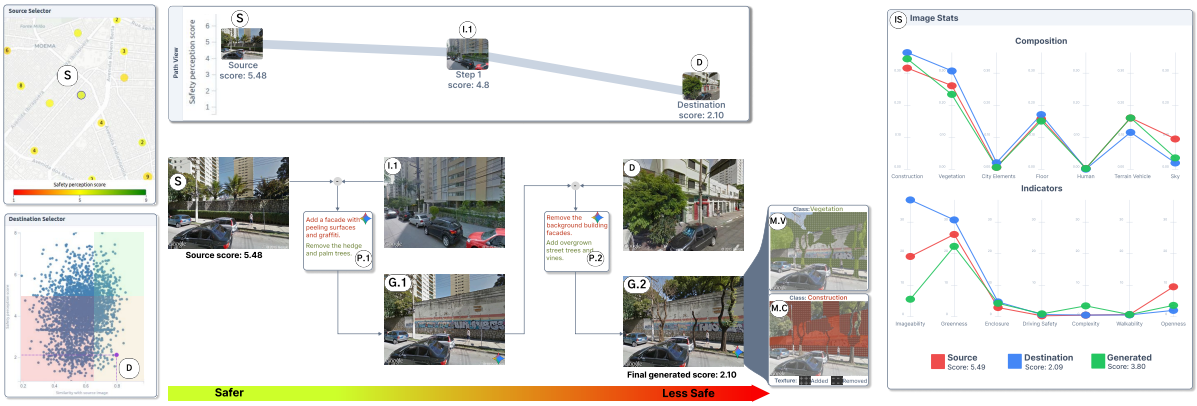

Figure 25 – Usage scenario showing a progressive decrease in perceived safety, transforming a relatively safe-looking image (S) into a less safe-looking one (G.2) through a two-step generation process. The method produces an intermediate result (G.1) that already appears more neglected. The most prominent change is the transformation of the decorated wall into a more deteriorated wall covered with graffiti. The Image Stats panel (IS) reflects this reduction in Vegetation and increase in Construction, which in turn decrease the Imageability and Greenness indicators, both often associated with perceived safety. Source: the author.

wall. The analyst selects as destination a visually similar image with a substantially lower perceived safety score (Figure 25D, Q-score: 2.10). The similarity graph reveals a short two-step path between source and destination, enabling a focused and interpretable reverse trajectory.

First step — introducing structural disorder. In the first editing step, the pipeline generates an intermediate result (Figure 25G.1) by transforming the source toward the intermediate reference image (Figure 25I.1). Because this reference scene is dominated by construction-related elements, the corresponding edit actions (Figure 25P.1) emphasize increasing the visual prominence of built structures while reducing vegetation. The most salient changes in the generated image are the introduction of a deteriorated wall with peeling paint and graffiti, and the removal of the hedge and palm trees that characterized the source scene. These modifications directly instantiate visual cues identified in Chapter 4 as associated with unsafe perception, and operationalize the disorder element vocabulary developed in Chapter 5.

Second step — reinforcing neglect. In the second step, the intermediate generated image is transformed further toward the destination (Figure 25G.2). The edit actions (Figure 25P.2) reduce the visibility of background facades and introduce overgrown trees and vines, reinforcing the impression of disorder and poor maintenance. Semantic change masks for the final step (Figure 25M.V and M.C) highlight the affected regions in the vegetation and construction categories, confirming that changes are localized and consistent with the stated editing actions.

Outcome analysis. The *Image Stats* view (Figure 25.IS) confirms the intended perceptual direction: compared with the source, the final generated image shows stronger construction prominence, reduced and disordered vegetation, and lower openness, with corresponding decreases in imageability, greenness, and openness. Applying the scoring model to the final generated image yields a Q-score of 3.80, representing a decrease of 1.68 points relative to the source and directionally consistent with the target score of 2.10.

This scenario demonstrates that the analytical value of UrbanCounterViz is not limited to generating improvements. The ability to explore transformations toward lower perceived safety supports diagnostic analysis, and reinforces the Broken Windows theory thread running through the thesis: the system can reproduce, in photorealistic form, the kinds of environmental deterioration that theory and prior research have long associated with reduced perceptions of safety.

7.2 Expert Feedback

To complement the usage scenarios with an external perspective on the system's utility and limitations, semi-structured interviews were conducted with three domain experts who had not been involved in the design or development of UrbanCounterViz. The goal was to gather qualitative feedback on three dimensions: usability (whether the system is clear and understandable), usefulness (whether it supports meaningful urban analysis), and interpretability (whether the outputs can be responsibly interpreted for design purposes). The expert interviews prioritize depth over breadth — seeking domain-grounded judgment about the realism of edits, the practical relevance of recommendations, and the ethical implications of deploying the system in planning contexts.

7.2.1 Interview Protocol

Three domain experts were recruited with experience in urban planning and urban analysis. The first is an architect with research experience in safety perception in public environments; the second is an academic researcher specializing in contemporary urban development; and the third is a specialist in urbanism. Together, these profiles provide complementary perspectives spanning the practical realities of urban design and intervention, the analytical and methodological dimensions of urban perception research, and the broader disciplinary context of urbanism as a field.

Three separate 45-minute semi-structured interviews were conducted, one with each expert. Each session followed the same structure: a brief introduction to the data collection process and the methodological pipeline underlying the system; a detailed presentation of the interface through representative examples and generated outputs from the usage scenarios; and an open discussion in which participants were asked to reflect on the

system's usability, usefulness, interpretability, and possible limitations. Participants were also invited to share suggestions for improvement and concerns about how the system's outputs might be interpreted or misused. To avoid bias, none of the experts is a co-author of this work, and none participated in the design process.

7.2.2 Findings

The following synthesizes the main themes that emerged across both interviews, organized by the three evaluative dimensions.

Usability.

All three experts described the overall system as clear and generally intuitive. The *Source Selector* and *Destination Selector* views were understood quickly, and the *Path View* was appreciated for making the transformation sequence explicit rather than presenting only a before-and-after comparison. The first expert noted that the textual descriptions in the *Generation View* should more closely reflect the actual modifications applied to the scene: in some cases, the stated editing actions described a broader or slightly different intervention than what was visually realized in the generated image, which could mislead analysts who rely on the text without examining the image carefully. The semantic change masks in the *Image View* were considered useful, though the same expert observed that they may be less necessary when the visual changes are already large and immediately apparent, and more valuable when differences are subtle and spatially localized. The second expert similarly described the overall pipeline as understandable and highlighted the perceived realism of the generated images as an important factor in building trust in the system's outputs. The third expert raised a concern about first-time users approaching the interface without sufficient context, and recommended that the system present an onboarding step in which representative examples of safe and unsafe images are shown alongside brief descriptions, allowing users to calibrate their understanding of the perception scale before beginning any analysis.

Usefulness. All three experts considered the system useful for analyzing how localized visual changes may affect perceived safety, though they offered nuanced assessments of which types of edits are most practically relevant. The first expert positively evaluated the use of contextual guidance through the similarity graph and discordant pairs, noting that grounding edits in differences observed between real scenes improves their plausibility compared with context-free generation. The second expert observed that some of the generated recommendations could plausibly inform practical urban interventions — particularly those involving vegetation, maintenance cues, and the removal of overhead cables — while others, such as roadway restructuring or the insertion of large built elements, would be difficult or costly to implement directly and might be better understood as directional signals rather than actionable proposals. All three experts emphasized that the most convincing

and useful edits are those that remain subtle and visually coherent within the scene; prominent additions such as large new buildings or foreground vehicles were perceived as less effective because they can distort the overall perception of the scene and draw attention away from the localized changes that are most analytically meaningful. The third expert additionally highlighted the *Image Stats* urban indicators panel as a particularly valuable component, noting that dimensions such as greenness, walkability, enclosure, and openness provide a meaningful and interpretable way to characterize the physical qualities of a street environment — one that goes beyond the perception score alone and connects the system's outputs to measurable aspects of urban space that practitioners already reason about in design and planning contexts.

Ethical considerations. Two of three experts raised concerns about how the system's outputs might be framed and interpreted in practice. The first recommendation is to clarify the meaning of the Q-score values displayed in the *Source Selector* map, to prevent users from interpreting them as objective measurements of actual safety rather than as estimates of perceived safety derived from crowdsourced pairwise judgments. This distinction — between perceived safety and objective safety — is not always immediately apparent to users unfamiliar with the Place Pulse methodology, and conflating the two could lead to misinterpretations with real policy implications. The third expert reinforced this concern from a different angle, recommending that the system explicitly contextualize users before they begin interacting with the interface — for instance, by presenting a curated set of example images spanning the safety score range, accompanied by short descriptions of the visual and semantic characteristics that distinguish them. Such an onboarding step would reduce the risk of users forming incorrect mental models of what the scores represent, and would help prevent unintended harm in cases where the system's outputs are interpreted as assessments of real or objective danger in specific locations. More broadly, the feedback across all three experts suggests that the practical value of the system depends not only on the visual quality and realism of the generated edits, but also on providing sufficient contextual explanation so that analysts can interpret the outputs and their implications responsibly.

7.2.3 Summary

Overall, the expert interviews indicate that UrbanCounterViz is clear, credible, and useful for examining how localized visual changes may influence perceived safety in urban environments. The most consistent finding across all three sessions is that the system's value is greatest when edits are short, visually subtle, and grounded in compositional differences between real scenes — conditions that correspond precisely to the design choices emphasized throughout this thesis. At the same time, the feedback identifies two important directions for improvement: strengthening the alignment between stated editing actions and generated visual outcomes, and providing clearer framing of the

system's outputs to prevent misinterpretation of perceived safety scores as objective safety measurements.

7.3 User Study

To complement the expert interviews with a broader and more structured assessment, a user study was conducted with participants drawn from a wider range of backgrounds. While the expert feedback in Section 7.2 provides domain-grounded qualitative judgment, the user study aims to assess how the system is perceived by the kinds of analysts, researchers, and practitioners who would encounter it in an applied context, and to generate quantitative usability evidence that can be compared against established benchmarks.

7.3.1 Study Design

A user study with 22 participants from diverse academic and professional backgrounds evaluated the analytical effectiveness and value-driven utility of UrbanCounterViz. The study intentionally recruited individuals without formal expertise in urban computing to assess whether the system makes complex visual behavior data accessible to a broader audience. For privacy reasons, no demographic information was collected from participants: neither nationality nor prior familiarity with São Paulo — the city from which the evaluated scenes are drawn — was recorded. This is a deliberate design choice, but also a limitation: because perceived safety is known to vary with cultural background and local familiarity (Ceccato, 2012; Ceccato and Loukaitou-Sideris, 2022), we cannot analyze whether, for example, participants acquainted with São Paulo judged the transformations differently from those seeing the scenes without local context. We return to this point when discussing the generalizability of the evaluation in Chapter 8. The evaluation design and questionnaire were inspired by the ICE-T framework Wall et al. (2018), which evaluates visualization systems across four dimensions: Insight, Confidence, Essence, and Time. Prior work using the ICE-T framework suggests that relatively small participant samples can provide meaningful feedback for exploratory evaluations of visualization systems Wall et al. (2018), supporting the use of this sample size. Although participants had prior experience evaluating interactive systems, none had previously used UrbanCounterViz.

7.3.2 Tasks and Procedure

The study was conducted both in person and online. Participants first received a 10-minute introduction to the UrbanCounterViz interface, covering the purpose and interactions of all six visual components: *Source Selector*, *Destination Selector*, *Path View*, *Image View*, *Generation View*, and *Image Stats*. During this introduction, participants

were also guided through the selection of a source and destination image, establishing a baseline understanding of scene similarity and safety score differences (T1.1) before proceeding to the structured tasks. Participants then completed four visual analytics tasks (U1–U4), designed to operationalize the high-level analytical tasks (T1–T3) defined in Section 6.3.2.

U1: Identify scene similarities and image composition differences required participants to first use the *Source Selector* and *Destination Selector* to identify and compare visually similar geo-located scenes based on their perceptual attributes and semantic composition (**T1.1**), and then use the *Path View* to analyze the transition sequence between the selected images and identify differences in object composition along the path (**T1.2**).

U2: Identify which components change the most required participants to use the *Image View*, *Path View*, and *Image Stats* to inspect and quantify semantic differences between the selected source image and its generated counterfactual, operationalizing **T2.1**, and to inspect how those differences accumulate across intermediate transition steps (**T2.2**).

U3: Analyze specific and minimal changes involved inspecting localized image composition through segmentation masks and textual descriptions in the *Generation View*, addressing **T3.1** by linking semantic editing actions with the corresponding highlighted regions in the *Image View* and relating observed modifications to shifts in perceptual scores.

U4: Compare visual features of original and generated images required participants to explore the full multi-step generation process using the *Generation View* and *Image Stats*. This task focused on relating urban indicators, semantic category distributions, perceptual scores, and similarity patterns across the original and generated images (**T3.2**), while revisiting semantic differences across generation steps (**T2.2**) to assess whether the cumulative transformation aligned with the intended perceptual direction.

7.3.3 Measures

For each task, participants were encouraged to think aloud Lewis (1982), explaining their reasoning and observations before answering both quantitative (QT) and qualitative (QL) questionnaires. In addition to the ICE-T items, participants were asked to judge the perception change between each original image and its modified counterpart — whether perceived safety increased, decreased, or remained unchanged — so that their assessment of the edit direction could be compared against the model's intended transformation. The complete form used during the user study, including the four tasks (U1–U4) and these perception-change questions, is provided in the appendix. Because the tasks emphasized open-ended exploration, visual reasoning, and hypothesis generation, traditional perfor-

mance metrics such as task completion time, completion rate, and error frequency were intentionally omitted, as these measures do not adequately capture the qualitative synthesis and insight validation central to value-driven evaluation frameworks. Both quantitative and qualitative results are reported below.

Quantitative Questions and Results

The quantitative questions were directly inspired by the ICE-T framework Wall et al. (2018), covering four dimensions: Insight, Confidence, Essence, and Time. Each dimension was assessed through two statements rated on a 7-point Likert scale (from "Totally Disagree" to "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).

Table 12 presents the ICE-T scores per participant across all four dimensions, where each score represents the average of the two corresponding Likert items. Overall, responses fall predominantly on the positive end of the scale, with minimal neutral or negative feedback. The dimension averages — Insight: 6.11, Confidence: 6.48, Essence: 6.23, Time: 6.36 — all exceed 6.0 on the 7-point scale, indicating consistently favorable evaluations across all four dimensions. Confidence received the highest average score, suggesting

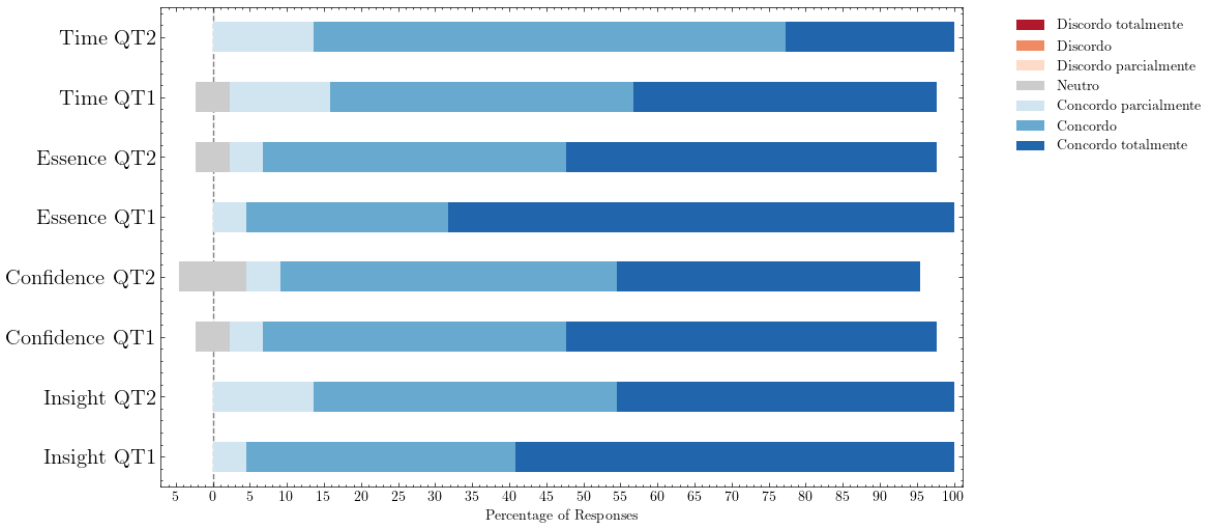

Figure 26 – Distribution of participant responses for questions QT1–QT8, grouped by ICE-T-inspired dimensions (Insight, Confidence, Essence, and Time). Bar length indicates the percentage of participants selecting each Likert level on a 7-point scale from "Totally Disagree" to "Totally Agree".

that participants found the coordination between visual components, composition values, and textual descriptions particularly effective for interpreting and verifying the system's outputs. Insight received the comparatively lowest average, with individual scores ranging more widely than in the other dimensions (from 5.0 to 7.0), pointing to some variability in how easily participants identified visually similar scenes with different safety scores during the initial selection stage. This variability aligns with the interface improvement suggestions reported in the qualitative findings below. Figure 26 further confirms the distribution of responses across the Likert scale.

Qualitative Questions and Findings

The qualitative questionnaire collected open-ended feedback on the usefulness of UrbanCounterViz and opportunities to improve the interface through two questions:

"*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).

Responses to QL1 and QL2 consistently identified the *Image View* and the *Generation View* as the most useful components for comparing images, identifying visual differences, and understanding the semantic descriptions associated with each generation step. Participants also emphasized that interactive features, particularly tooltips, helped reduce visual clutter while preserving access to contextual information.

Regarding the data panels, participants reported that the *Image Stats* semantic

Table 12 – ICE-T evaluation scores per participant across the four dimensions. Each score is the average of two Likert items (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**

composition panel was intuitive for understanding changes in detected objects, whereas the urban indicators panel required greater cognitive effort because its interpretation depended on numerical values rather than direct visual cues. This finding suggests that visual object representations are more immediately interpretable than quantitative indicator summaries, and points to a potential redesign direction for the urban indicators panel — for instance, through icon-based or color-coded encodings that reduce reliance on numerical reading.

Suggestions for improvement primarily focused on increasing user control over the editing process. Rather than only inspecting the system-generated textual actions and descriptions, participants expressed interest in proposing and evaluating their own modifications interactively. Additional recommendations included allowing panels to be resized and increasing the prominence of the map view to facilitate the initial selection of source and destination locations, consistent with the lower Insight scores observed in the quantitative results.

7.3.4 Summary

The user study provides structured quantitative and qualitative evidence that UrbanCounterViz is effective and accessible to a broad audience of analysts and researchers without specialized expertise in urban computing. ICE-T scores across all four dimensions exceeded 6.0 on a 7-point scale — a range that indicates positive but not uncritical evaluations, consistent with participants identifying both strengths and concrete areas for improvement. These scores reflect a system that is effective and accessible while retaining room for refinement, particularly in the scene selection workflow and urban indicators panel. Confidence received the highest average score (6.48), corroborating the expert feedback finding that the coordination between visual components and textual descriptions is the system's strongest usability asset. The *Image View* and *Generation View* were most frequently cited as the most valuable components, reinforcing the design choice to make the editing process explicit and inspectable at each step rather than presenting only a final result.

At the same time, the study identifies clear directions for improvement. Insight received the comparatively lowest average score and the widest spread of individual responses, with qualitative suggestions about map prominence and panel resizability pointing to friction in the initial scene selection stage — the entry point to the analytical workflow. The urban indicators panel, while appreciated in principle by the domain experts, required more cognitive effort than the semantic composition panel among the broader participant sample, suggesting that its visual encoding could be made more immediate. Finally, the recurring request for interactive editing control — the ability to propose and test one's own modifications rather than only inspect system-generated ones — aligns directly with the future work direction identified in Chapter 8: enabling user-steerable editing actions as a next development step for UrbanCounterViz.

7.4 Final Considerations

This chapter presented the evaluation of UrbanCounterViz through three complementary methods: usage scenarios, expert feedback, and a user study. Taken together, they assess the system from different but mutually reinforcing angles — demonstrating its analytical range through structured scenarios, validating its practical credibility through domain expert judgment, and measuring its usability and requirement coverage through a broader participant sample.

Several findings deserve particular emphasis as they connect the evaluation back to the thesis contributions as a whole. First, the grounding strategy — deriving edits from differences between real adjacent scenes rather than from unconstrained generation — was consistently identified across all three evaluation methods as the property most responsible

for the system's interpretability and credibility. Second, the explicit transparency of the pipeline — the ability to inspect semantic differences, editing actions, generated images, and multi-dimensional outcome summaries at every step — was recognized as essential for the human-in-the-loop workflow to function as intended. Third, the bidirectional capability demonstrated in Scenario 3 extends the system beyond improvement planning to diagnostic analysis, enabling a more complete understanding of the visual factors that shape safety perception in both directions. Fourth, the user study provides quantitative grounding for the qualitative observations of the expert interviews, providing preliminary evidence about whether the usability and requirement coverage perceived by domain practitioners also holds for a broader user population.

Chapter 8 synthesizes these findings with the broader contributions of the thesis. It situates the evaluation results within the four-contribution framework established in Chapter 1 — the systematic review, the visual element analysis, the UrbanCFX pipeline (Chapter 5), and the UrbanCounterViz system with its reusable interpretable workflow — and reflects on the extent to which each contribution addresses the methodological gaps identified in Chapter 3. Chapter 9 then draws final conclusions, acknowledges the limitations of the work as a whole, and outlines directions for future research.

8 Discussion

The four contributions of this thesis — the systematic literature review (Chapter 3), the investigation of visual elements in safety perception through UrbanFormer (Chapter 4), the UrbanCFX pipeline and UrbanPD4k dataset (Chapter 5), and the UrbanCounterViz system with its reusable interpretable workflow (Chapters 6–7) — each address a distinct but related aspect of the challenge of interpretable urban safety perception analysis. Taken individually, each makes a meaningful contribution: the systematic review maps the methodological landscape of the field and identifies the gaps this thesis addresses; UrbanFormer identifies which visual elements drive model attention during safety prediction; UrbanCFX translates feature-space counterfactuals into natural language recommendations; and UrbanCounterViz generates photorealistic, grounded, and interactively explorable urban scene transformations. However, the most important insights emerge from considering these contributions together, as stages in a cumulative research argument about what it means to understand, explain, and act on computational estimates of urban safety perception.

This chapter synthesizes those insights. Section 8.1 integrates the main findings across all contributions and identifies the overarching lessons they offer for the field of urban perception research. Section 8.2 discusses the limitations of the full research pipeline, going beyond the chapter-level limitations already presented in each contribution. Section 8.3 outlines future research directions that follow directly from the work presented here.

8.1 Synthesis of Findings

8.1.1 Grounding Counterfactual Reasoning in Real Urban Data

A thread that runs through all three empirical contributions is the principle that interpretable analysis of urban perception should be grounded in differences observed between real scenes rather than derived from model optimization or unconstrained generation. This principle was established theoretically in Section 6.4.3 through the formal concept of discordant pairs — while the intuition behind comparing similar scenes with divergent scores already motivated the analyses of Chapters 4 and 5 — and realized as a fully interactive, path-guided system in UrbanCounterViz.

The evaluation results in Chapter 7 provide concrete evidence for why grounding matters. In the usage scenarios, edits derived from real discordant pairs produced more visually coherent and analytically interpretable transformations than longer trajectories in which the grounding constraint weakened as path length increased. The domain experts

corroborated this observation, independently identifying contextual grounding as the property most responsible for the system's practical credibility. This convergence between quantitative scenario analysis and qualitative expert judgment strengthens the argument that the grounding strategy is not merely a methodological choice but a substantive contributor to the system's value.

From a broader perspective, this finding connects to a well-documented limitation of counterfactual explanation methods in machine learning: unconstrained counterfactuals often produce inputs that are optimal for shifting a model's prediction but unrealistic or uninterpretable as real-world interventions (Wachter et al., 2017; Mothilal et al., 2020). The approach proposed in this thesis addresses this limitation in the urban perception domain by replacing synthetic counterfactuals with observation-grounded ones, using the structure of a visual similarity graph to constrain transformations to the space of changes that actually occur between real urban scenes.

8.1.2 The Role of Semantic Representation

Across all three empirical contributions, semantic scene representation played a pivotal role — but at different levels of granularity and for different purposes. In Chapter 4, the 150-category ADE20K pixel-ratio vector enabled the identification of which broad object categories (trees, sidewalks, walls, bare ground) were most associated with safe and unsafe predictions. In Chapter 5, this vocabulary was extended with 13 manually annotated disorder elements (graffiti, overhead cables, broken sidewalks, damaged walls) that are invisible to standard segmentation models but highly relevant to human safety judgments. In UrbanCounterViz (Chapters 6–7), the two vocabularies are combined: ADE20K categories provide the compositional backbone for similarity graph construction, path guidance, and semantic difference computation, while the disorder element vocabulary from UrbanPD4k informs the editing action generation stage and enriches the semantic summaries displayed in the interface.

This progressive enrichment of the semantic vocabulary reflects a genuine analytical insight: the visual cues that most strongly differentiate safe from unsafe urban environments are often *within-category* distinctions that standard segmentation taxonomies conflate. A maintained wall and a graffiti-covered wall occupy the same ADE20K category, yet they produce very different safety judgments. A sidewalk lined with trees and a bare dirt path both register as combinations of *floor* and *vegetation*, yet they signal very different levels of environmental order. The disorder annotation effort in Chapter 5 and its integration into UrbanCounterViz represent a methodological response to this limitation, and the usage scenarios in Chapter 7 demonstrate the consequence: the most informative and analytically credible edits in Scenarios 1 and 3 involve precisely these within-category transformations — the introduction of a graffiti-covered deteriorated wall, the replacement

of organized hedges with overgrown vegetation.

8.1.3 Transparency and Human Oversight as System Properties

One of the consistent findings across the evaluation is that transparency — the ability to understand not only what the system produces but why — is not a peripheral feature but a central determinant of the system's usefulness. The domain experts in Chapter 7 did not evaluate UrbanCounterViz primarily on the visual quality of its generated images but on whether they could understand and trust the reasoning behind them. This observation has important implications for the design of analytical systems in high-stakes domains.

Prior visual analytics systems for urban imagery, such as StreetVizor (Shen et al., 2017) and Urban Mosaic (Miranda et al., 2020), prioritized exploration and retrieval over explanation. Prior generative methods, such as FaceLift (Joglekar et al., 2020) and URSimulator (Hu et al., 2025), prioritized visual quality over interpretability of the transformation rationale. UrbanCounterViz is designed from the outset to make every step of its pipeline inspectable: the similarity graph structure, the discordant pair selection, the semantic difference computation, the editing action generation, the photorealistic output, and the multi-dimensional outcome summary are all exposed to the analyst through coordinated views. The evaluation results suggest that this transparency is not merely a design virtue but a functional prerequisite for the system to be considered credible and useful by domain practitioners.

The system also reinforces a broader point about the relationship between visual analytics and machine learning interpretability. The VIS4ML framework (Sacha et al., 2019) and related work on explainable AI (Choo and Liu, 2018; La Rosa et al., 2023) have argued that human-in-the-loop interaction is essential for responsible deployment of ML models in consequential domains. The design and evaluation of UrbanCounterViz provide a concrete case study in which this principle is applied to the domain of urban perception: the human analyst is not a passive consumer of model outputs but an active participant who selects paths, interprets intermediate steps, and decides when a generated transformation is analytically credible.

8.1.4 Bidirectionality as a Diagnostic Tool

Scenario 3 in Chapter 7 demonstrated a dimension of UrbanCounterViz's analytical value that was not explicitly anticipated in the design requirements but emerged naturally from the system's bidirectional capability: the ability to generate transformations toward *lower* perceived safety as a means of diagnosing which visual cues most strongly signal disorder. This direction of analysis is complementary to the improvement-oriented

scenarios and connects the system to a long tradition in urban research of understanding deterioration and disorder as active processes rather than merely as the absence of order.

From a Broken Windows theory perspective (Wilson and Kelling, 1982), which motivates the thesis from its opening chapter, the ability to generate and inspect deterioration scenarios is as analytically important as the ability to generate improvements. Understanding why a scene is perceived as unsafe requires understanding not only what elements are associated with safety (trees, maintained sidewalks, organized built environment) but also what elements actively signal disorder (graffiti, overgrowth, deteriorated surfaces). Scenario 3 shows that UrbanCounterViz can generate these disorder signals in photorealistic form, grounded in real discordant pairs, and trace their semantic sources through the coordinated interface views. This gives researchers a tool for exploring and formulating hypotheses about visual disorder — for example, whether graffiti alone appears sufficient to shift a scene from safe to unsafe in a generated transformation, or whether it tends to require the co-occurrence of vegetation disorder and structural deterioration.

8.2 Limitations

The following limitations apply to the full research pipeline presented in this thesis, extending beyond the chapter-level limitations already discussed in Sections 4.4.4 and 5.5, and the evaluation limitations identified in Section 7.4.

Generalizability of findings. All empirical contributions rely primarily on Place Pulse 2.0, a dataset collected through online crowdsourcing from a global but demographically non-representative population. The safety perceptions captured in this dataset reflect the aggregate judgments of a specific annotator population at a specific point in time, and may not generalize across cultural contexts, demographic groups, or cities with different visual and infrastructural characteristics. The UrbanCFX analysis in Chapter 5 further restricts its scope to Rio de Janeiro, whose socio-spatial characteristics — particularly the co-occurrence of informal settlements, visible disorder elements, and stark contrasts in infrastructure quality — may not be representative of other urban contexts. The usage scenarios in Chapter 7 use São Paulo, which shares some of these characteristics, but systematic cross-city validation remains an important open question.

Perception versus objective safety. Throughout this thesis, the term "safety perception" is used consistently and carefully to refer to human perceptual judgments derived from visual information, not to actual crime rates, physical hazards, or other objective measures of safety. However, as the expert feedback in Section 7.2 highlighted, this distinction is easily blurred in practice, particularly when perceptual scores are displayed on a map that superficially resembles a crime or risk map. Future deployments of UrbanCounterViz should invest in interface-level communication of this distinction — for example, through

explicit labeling, contextual tooltips, and onboarding materials — to ensure that outputs are interpreted responsibly by planners and policymakers who may not be familiar with the crowdsourced perception methodology.

Generation latency. Each editing step in the two-stage VLM pipeline requires approximately 18 seconds to complete, primarily due to the inference time of the Gemini image-editing model. For short trajectories (two to three steps), this is manageable within a single analytical session. For longer trajectories (five to six steps), the accumulated latency approaches two minutes, which interrupts the analytical flow and limits the number of alternative paths an analyst can explore in a reasonable time frame. This constraint is particularly relevant for the human-in-the-loop use case that UrbanCounterViz is designed to support: if the cost of exploring a single alternative trajectory is several minutes of waiting, analysts may converge prematurely on the first trajectory explored rather than comparing multiple options.

Dependence on proprietary generation models. A related limitation concerns the reliance on a proprietary, closed-source generation model. The two-stage editing pipeline uses Gemini models accessed through a commercial API, which constrains reproducibility: model versions may change or be deprecated, inference is not fully deterministic, and the internal generation process cannot be independently audited. While the grounding strategy and the surrounding pipeline are model-agnostic by design, replicating the specific visual results reported in this thesis requires access to the same model version used during evaluation.

Spatial control in image editing. The VLM image-editing model exhibits limited spatial precision over specified semantic regions. In several instances across the usage scenarios — particularly in longer trajectories — the model introduced large foreground elements or globally restructured scene geometry in ways not implied by the semantic difference vector. Segmentation-based masks help analysts identify these unintended changes after the fact, but they do not prevent them during generation. Future work integrating spatially conditioned inpainting or mask-guided diffusion models could address this limitation by constraining edits to specific image regions rather than allowing the generative model to modify the scene globally.

Annotation scalability. The 13-category disorder annotation effort in Chapter 5 required two trained annotators — both PhD students with prior experience in urban perception and long-term residence in Rio de Janeiro — working on 3,659 images, with reconciliation to resolve disagreements. While this effort substantially enriched the semantic vocabulary available to the system, it is not straightforwardly scalable to larger image collections or to new cities where the relevant disorder elements may differ. Automated disorder detection models trained on the UrbanPD4k dataset could address this bottleneck, but would require careful validation to ensure that domain-specific characteristics — such as

the visual appearance of informal housing in São Paulo versus Rio de Janeiro versus other cities — are adequately captured.

Static imagery and the neglect of temporal and weather variation. A limitation that cuts across every contribution of this thesis is that all analyses rest on single, static street-view snapshots. Each location is represented by one image, captured at one moment, under one set of weather and lighting conditions. This has a direct and easily overlooked consequence for the results: in Chapter 4, the sky emerged as a high-prevalence but near-neutral category, and it would be tempting to conclude that the sky simply does not matter for perceived safety. That conclusion would be an artifact of the representation, not a finding about perception. Our features encode only *how much* sky is visible, never its *state* — overcast versus clear, bright daylight versus dusk, dry versus rain — because those factors are held fixed by the single-snapshot design and discarded by the pixel-ratio vector. Environmental criminology has shown precisely the opposite of neutrality for these factors: perceived safety, fear of crime, and even the spatial–temporal distribution of violent crime vary systematically with weather, season, and time of day, including evidence from São Paulo that homicide patterns track weather and temporal variation, and broader work establishing how the temporal rhythm of the environment shapes fear and the perception of urban space (Ceccato, 2005; Wikström et al., 2010; Ceccato and Uittenbogaard, 2014; Ceccato and Loukaitou-Sideris, 2022). The apparent neutrality of the sky in this thesis should therefore be read narrowly, as a statement about visible sky *area* in a fixed-condition dataset, and not as a claim that sky conditions are perceptually unimportant. Capturing these effects is an important direction for future work and would require temporally sampled imagery (e.g., the same location across times of day, seasons, and weather) together with appearance-aware features rather than presence-only descriptors.

Presence-weighted relevance and the under-crediting of small elements. The element-level relevance measure introduced in Chapter 4 — the overlap between model attention and each segmentation mask, averaged over the dataset — is inherently *presence-weighted*, and this deserves more emphasis than its chapter-level treatment gives it. Because relevance is quantified as spatial area overlap and then averaged across all images, large and ubiquitous categories such as trees or buildings accumulate high scores partly by virtue of the sheer image area they occupy, while a small element that covers only about 1% of a scene — graffiti, a broken sign, a garbage bag — is pushed toward a near-zero average impact even in the images where the model attends to it strongly. The metric, in other words, rewards how *much* of the image an element fills rather than how *salient* it is where it appears, which risks systematically under-crediting exactly the compact disorder cues that the Broken Windows framing places at the center of perceived safety. This is not merely a scaling nuisance: it means the relevance ranking should not be read as a direct ranking of perceptual importance for small objects. Addressing it would require decoupling occurrence frequency from per-object salience — for instance, normalizing overlap by

each element's own area, conditioning the average only on images where the element is present, or complementing area-based IoU with attention-density measures that reward concentrated activation over a small region. We regard this as a concrete methodological refinement for future explanation work rather than a settled aspect of the current pipeline.

Graffiti as a single category: disorder versus art. Throughout the disorder vocabulary of this thesis, graffiti is treated as one category and, implicitly, as a signal of neglect. This warrants stronger qualification than the annotation chapter alone provides. Graffiti is not a homogeneous cue: illicit tagging and vandalism sit at one end, but muralism, commissioned public art, and sanctioned street art sit at the other, and the latter are frequently perceived not as disorder at all but as evidence of investment, cultural identity, and an active, cared-for public space — and may therefore raise, rather than lower, perceived safety. Because the present work collapses these cases into a single annotated class, and because the counterfactual and editing stages inherit that class, the system may treat the removal of any graffiti as an unambiguous improvement, when in some contexts the opposite could hold. This is a genuine conceptual limitation of the disorder taxonomy, not only an annotation-cost issue: a more faithful treatment would sub-categorize graffiti by type (e.g., tags versus murals) and study whether artistic graffiti behaves as a distinct, potentially safety-increasing element. We leave this refinement of the disorder vocabulary to future work.

8.3 Future Work

The limitations identified above suggest several concrete directions for future research, organized from the most immediate technical extensions to broader methodological and application directions.

User-steerable editing actions. The most immediate extension of UrbanCounterViz is to allow analysts to refine or override the editing actions generated in Stage 1 of the pipeline before they are passed to the image-editing model. Currently, the *Generation View* displays the actions for inspection, but the analyst cannot modify them. A chat-like interface in which analysts could adjust prompt wording, exclude specific categories, or modulate the magnitude of proposed changes would transform generation into a more genuinely collaborative process and reduce the gap between intended and realized edits identified in the expert feedback.

Free-form counterfactual manipulation. Beyond steering the existing pipeline, future work could enable analysts to specify desired semantic compositions directly — for example, by adjusting sliders in the *Image Stats* view that represent target proportions for each semantic category — rather than navigating only a precomputed similarity graph. This capability would expand the space of explorable transformations while preserving the

system's analytical framing, and would complement the path-guided approach by enabling hypothesis-driven exploration in addition to data-driven discovery.

Cross-city and local-context adaptation. As noted in Chapter 5, the restriction to a single city in the UrbanCFX analysis was a deliberate methodological choice to eliminate inter-city confounds and ensure a controlled investigation of disorder elements. Extending this approach across multiple cities nevertheless remains an important open direction. The current system is instantiated on Place Pulse 2.0, which provides broad geographic coverage but limited local specificity. Extending the system to incorporate city- or district-specific perception datasets — collected through resident surveys, local crowdsourcing platforms, or mobile studies — would ground its recommendations in the safety judgments of the communities most directly affected by the urban environments being analyzed. This direction is particularly important for ensuring that UrbanCounterViz's outputs are responsive to culturally specific visual grammars and locally meaningful disorder cues, rather than reflecting the aggregate judgments of a globally distributed but demographically unrepresentative annotator pool.

Integration with urban design workflows. The usage scenarios in Chapter 7 and the expert feedback in Section 7.2 suggest that UrbanCounterViz's outputs — particularly the editing actions and generated images from short, coherent trajectories — could function as inputs to urban design processes rather than as standalone analytical results. Future work could explore how the system's recommendations could be exported in formats compatible with urban design tools (e.g., annotated street-level views, semantic change reports, or prioritized intervention lists), and could evaluate whether UrbanCounterViz-informed interventions improve perceived safety in longitudinal field studies.

Evaluation with broader participant samples. The expert feedback study in Chapter 7 involved three participants, which is appropriate for a qualitative, interview-based investigation but does not provide the statistical power needed to generalize conclusions about usability or usefulness. Future work should include more extensive user studies with participants spanning a broader range of domain expertise, from professional urban planners and architects to community residents and local government officials. This would provide a more complete picture of how different user populations interpret and act on UrbanCounterViz's outputs.

8.4 Final Considerations

The central claim of this thesis is that understanding urban safety perception requires more than accurate predictive models: it requires tools that support interpretable, grounded, and planning-oriented reasoning about the visual factors that drive human judgments of urban environments. The evidence gathered across Chapters 3 through 7

— from the gaps surfaced in the literature, to the elements identified by UrbanFormer (Chapter 4), to the plain-language recommendations of UrbanCFX (Chapter 5), to the inspectable transformations produced by UrbanCounterViz (Chapters 6–7) — speaks to this claim, even if a single dissertation can only take it so far.

The synthesis presented in this chapter suggests three broader lessons for the field. First, grounding counterfactual reasoning in real data — through the discordant pair framework and similarity graph structure — is not merely a methodological preference but a substantive contributor to the interpretability and credibility of the resulting recommendations. Second, semantic vocabulary matters: the within-category distinctions that standard segmentation taxonomies miss (maintained versus deteriorated walls, organized versus overgrown vegetation) are precisely the cues that most strongly differentiate safe from unsafe urban scenes in human perception. Third, transparency and human oversight are not peripheral features of analytical systems for high-stakes domains but central determinants of their usefulness, as demonstrated by the consistent emphasis on pipeline inspectability in both the usage scenarios and the expert feedback.

Chapter 9 concludes the thesis by revisiting the four specific objectives established in Chapter 1, summarizing the contributions against each objective, and reflecting on the broader implications of this work for the study of urban perception, the design of interpretable AI systems for the built environment, and the role of visual analytics in connecting computational models with planning-oriented domain knowledge.

9 Conclusions

This dissertation set out to understand what makes a street feel safe or unsafe, and to ask whether that understanding could be turned into something more useful than a prediction. The chapters that followed pursued that question in three steps, mapped out by the systematic review that opened the work: moving from identifying visual elements, to describing plausible changes to them in plain language, to finally making those changes visible and inspectable. What the evidence gathered along the way supports is that this is possible, and that each step is genuinely necessary: none of the three alone would have been enough.

9.1 Contributions Revisited

Each of the four specific objectives set out in Chapter 1 — mapping the state of urban perception research, identifying the visual elements that shape perceived safety, translating these associations into human-readable recommendations, and making plausible changes visible and inspectable — is met by the corresponding contribution revisited below.

The systematic review in Chapter 3 screened 2,307 papers and synthesized 234 studies, mapping the methodological landscape of urban perception research using SVI and identifying four persistent gaps — limited insight into visual factors, static explanations, unrealistic generative edits, and the absence of interactive human-in-the-loop tools — that directly motivated all subsequent contributions.

Chapter 4 addressed the first gap through UrbanFormer, which combined CNN representations with OneFormer pixel-ratio vectors and applied Grad-CAM analysis across more than 108,000 images to identify which visual elements — trees, sidewalks, walls, bare ground — most strongly shape perceived-safety predictions. Chapter 5 deepened this with UrbanCFX and the UrbanPD4k dataset, introducing 13 manually annotated physical disorder elements and a pipeline that translates counterfactual difference vectors into plain-language design recommendations via a LLM, closing the communicative gap between model outputs and urban practitioners.

Chapters 6 and 7 introduced UrbanCounterViz, a visual analytics system that extends counterfactual reasoning into the image domain. By grounding every edit in differences observed between real discordant scene pairs, deriving photorealistic transformations via a two-stage VLM pipeline, and exposing the full process through a coordinated six-view interface, UrbanCounterViz makes grounded, inspectable, and steerable urban scene exploration possible. Evaluation through usage scenarios, expert interviews, and a

22-participant user study provided converging evidence of the system's analytical value, with ICE-T scores exceeding 6.0 across all four dimensions and domain experts consistently identifying grounding and transparency as the properties most responsible for the system's credibility.

Alongside UrbanCounterViz itself, this thesis's fourth contribution also includes a reusable interpretable workflow — independent of any specific system implementation — that formalizes how model predictions, semantic scene differences, and grounded visual transformations can be systematically connected. This workflow, anchored in the discordant pair concept and the visual similarity graph, provides a generalizable template for future work on interpretable urban perception analysis.

9.2 Final Remarks

Urban environments shape behavior and well-being in consequential ways. The ability to reason computationally about which visual characteristics signal perceived safety or disorder — and to explore how plausible changes to those characteristics might shift human judgment — is therefore not merely a technical achievement but a practical contribution to urban design and public policy.

Taken together, these contributions reframe the analysis of perceived urban safety as more than a predictive modeling task. They show how data-driven perception models, semantic scene representations, counterfactual reasoning, generative image editing, and visual analytics can be connected into a single workflow for grounded, interpretable, and human-inspectable reasoning about urban visual environments.

The framework established in this thesis — grounding counterfactual analysis in real observed differences, making every pipeline stage transparent and inspectable, and connecting model behavior to planning-oriented recommendations — provides a principled foundation for future extensions: cross-city generalization, user-steerable editing, and integration with professional design workflows. These remain as open directions; the foundation on which they can be built is the contribution of this work.

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.