ISPRS Journal of Photogrammetry and Remote Sensing 236 (2026) 622–639

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Bridging street view coverage disparities through geographic identity preserving generation from satellite view

Zongrong Li [a,b], Fan Zhang [c], Shaoqing Dai [d,e,*], Wufan Zhao [a,*]

[a] Thrust of Urban Governance and Design, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[b] Spatial Sciences Institute, University of Southern California, Los Angeles, United States
[c] Institute of Remote Sensing and Geographic Information System, School of Earth and Space Sciences, Peking University, Beijing, China
[d] School of Resource and Environmental Sciences, Wuhan University, Wuhan, China
[e] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences

ARTICLE INFO

Keywords:
Cross-view image generation
Satellite-to-street view synthesis
Geographic identity
Urban data equity

ABSTRACT

Street view imagery (SVI) provides a human-centric view of urban environments and is widely used for analyzing greenery, mobility, socioeconomics conditions, health outcomes, and safety perception. However, its coverage is highly uneven, with developing regions systematically underrepresented, limiting the inclusiveness of SVI-based analytics. To address this, we propose GeoIdentity-Sat2Street, a geographic identity preserving framework, that leverages satellite imagery to expand SVI coverage. Our framework first applies a polar-transformation conditional GAN to synthesize plausible street view perspectives, then refines them with a diffusion-based generator conditioned on semantic captions, location metadata, and structural priors. This design enforces geometric consistency while explicitly preserving geographic identity, an overlooked but critical aspect of urban distinctiveness. We conduct a comprehensive evaluation against baselines, covering generation quality, geographic identity preservation, and global scalability. First, on classic cross-view benchmarks (CVUSA and CVACT), our method achieves the highest SSIM (0.427/0.537), lowest LPIPS (0.345/0.317), and best mIoU (0.054). Next, to assess geographic identity fidelity, we introduce MultiCities Dataset, a benchmark of 50,000 paired satellite–street-view images across five cities on five continents. Our method achieves the highest Silhouette Score (0.222), lowest inter-city variance (0.031), and clearly separated clusters in t-SNE, demonstrating superior preservation of regional visual identity; GPT-based evaluation further confirms realism and semantic alignment (score 10.587–10.824). Finally, we apply our model to Kathmandu, Nepal achieving about a 28% improvement in usable street-view coverage, from 72% to dense roadside coverage. Overall, this work highlights the proposed identity preserving two-step pipeline as central to equitable SVI generation and provides a scalable framework toward globally representative urban analytics.

1. Introduction

The rapid growth of urban data has enabled unprecedented opportunities to study cities at scale. Among them, street view imagery (SVI) is unique in offering an irreplaceable ground-level and human-centric view of urban environments, revealing pedestrian-experience features such as street view greenery, façade details, and safety-related cues that cannot be observed from satellite imagery. This perspective has made SVI a crucial data source for studying greenery, mobility, socioeconomic conditions, health outcomes, and safety perceptions (Biljecki and Ito, 2021; Zhang et al., 2024). However, the application of SVI in urban research is strongly constrained by its uneven global coverage.

A recent systematic review showed that more than half of published studies were conducted in the United States, with only a handful in China, Brazil, and other developing countries (Dai et al., 2024). Coverage analyses confirm this imbalance, reporting substantial gaps in façades and streetscapes, with developing regions systematically underrepresented (Fan et al., 2025; Alpherts et al., 2025). Empirical work in Africa further shows that Google Street View coverage is biased toward central or tourist areas, leaving many neighborhoods undocumented (Umar et al., 2023). More generally, reviews highlight that SVI observations remain concentrated in Europe and North America, with

* Corresponding authors.
E-mail addresses: shaoqing.dai@outlook.com (S. Dai), wufanzhao@hkust-gz.edu.cn (W. Zhao).

https://doi.org/10.1016/j.isprsjprs.2026.03.049
Received 1 December 2025; Received in revised form 15 March 2026; Accepted 30 March 2026

limited representation across the Global South (Yin et al., 2023). Visualization (Fig. 1) of global coverage illustrates the same disparities: while cities in developed countries often exhibit near-complete representation, large gaps persist across Asia, Africa, and South America. Such uneven coverage not only constrains the scope of SVI-based research but also risk introducing systematic biases in understanding urban environments. For instance, Wang et al. (2025) showed that differences in SVI platform coverage produced significant variance in greenery and built-environment measurements, demonstrating how uneven SVI distribution may distort urban analysis outcomes. This motivates the need for scalable approaches to supplement SVI coverage in data-scarce regions.

To mitigate these disparities, researchers have explored generating or inferring street view information from alternative data sources, including soundscapes (Wang et al., 2025), drone imagery (Sun et al., 2023) and satellite view imagery (Li et al., 2024; Shi et al., 2022). Among these, satellite view imagery has emerged as the mainstream modality, due to its global availability, consistency, and relatively uniform quality. Building on this foundation, studies have advanced satellite-to-street view generation through two complementary directions: strengthening geometric correspondence and improving generative modeling capacity. GAN-based methods demonstrated the feasibility of viewpoint transformation, while geometry-aware techniques such as height-based and differentiable projection further improved alignment (Lu et al., 2020; Shi et al., 2020). Diffusion models boost appearance and semantic modeling, with CrossViewDiff and Satellite to GroundScape leveraging cross-view attention and structural priors for richer conditional generation (Xu and Qin, 2025; Li et al., 2024). While advances in generative techniques have improved perceptual realism and geometric fidelity, these efforts largely focus on generic visual quality rather than region-specific appearance.

Yet these works were not explicitly motivated by the challenge of bridging SVI coverage gaps. Their primary aim has been to advance the technical problem of cross-view generation, with relatively little attention to geographic identity, a core dimension of urban identity and distinctiveness shaped by materials, vegetation, signage, and everyday details. As Guo et al. (2025) emphasize, the uniqueness of urban form arises from both identity and distinctiveness, often captured in ordinary streets rather than iconic landmarks. Ignoring these identity-related dimensions risks producing homogenized views that fail to reflect local character, undermining the usefulness of generated imagery for downstream applications such as safety perception, greenery assessment, socio-spatial inequality analysis, and tourism. Ensuring geographic identity is vital for globally relevant SVI generation.

To this end, we develop GeoIdentity-Sat2Street, a geographic identity preserving street view generation framework designed to supplement coverage in regions where street view imagery is sparse. Our pipeline follows a two-step design: a polar-transformation-based conditional GAN first converts satellite imagery into street view perspectives, narrowing the geometric gap between satellite and street views; a diffusion-based generator then refines these views by integrating semantic captions, location metadata, and structural priors to preserve geographic identity. By jointly ensuring structural accuracy and stylistic fidelity, our method advances cross-view generation toward globally meaningful applications.

Building on this design, our study makes three contributions. First, we propose a identity preserving two-step pipeline that generates realistic street view imagery from satellite data while maintaining geographic consistency, substantially improving realism and regional fidelity. Second, we construct MultiCities Dataset, a benchmark of 50,000 paired satellite and street-level images from five cities on five continents, enabling systematic evaluation of both structural quality and stylistic preservation. Third, we further validate the framework's generalization in Kathmandu, Nepal, where the generated imagery supplements missing street view data and enables downstream road-safety perception analysis. These results confirm that our method is capable of producing geographically faithful street views at a large scale, contributing to more equitable and comprehensive urban research worldwide.

2. Related work

2.1. Street view imagery generation

Street view imagery generation from satellite imagery has advanced through two key dimensions: the modeling of geometric relationships

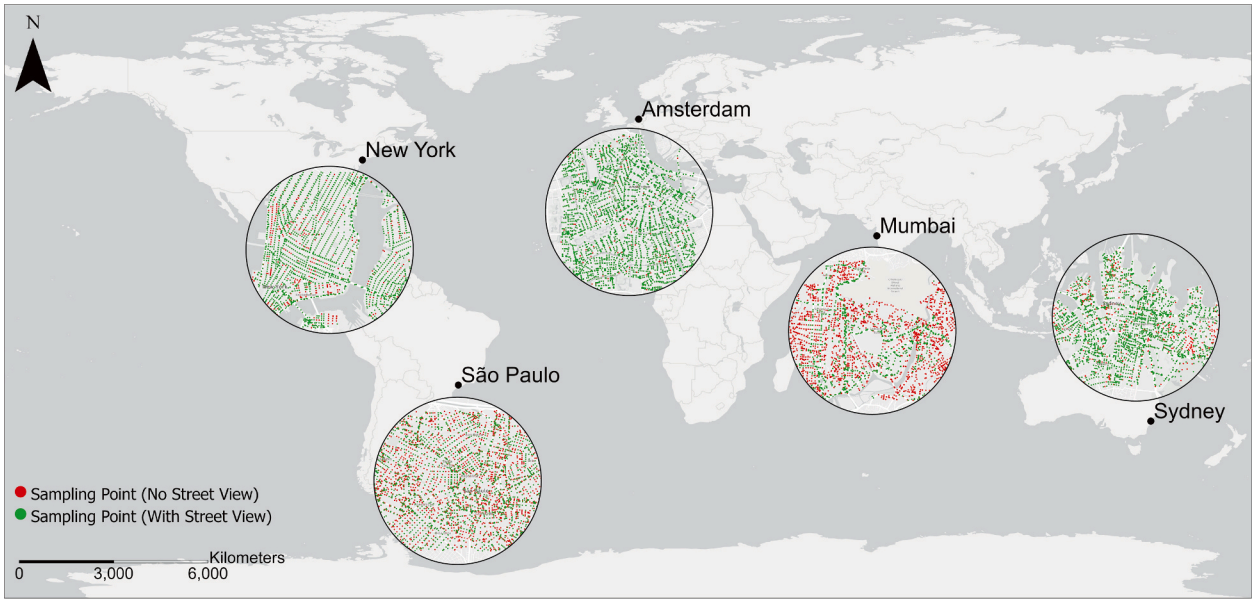

Fig. 1. Street view imagery coverage in selected cities. *Note: Street view coverage differs widely between developed and developing cities: Amsterdam (96.1%), New York (85.3%), and Sydney (82.5%) show high availability, whereas São Paulo (52.1%) and Mumbai (35.5%) remain far less covered; Green points indicate sampling locations whose 250 m query radius contains street view imagery, whereas red points indicate locations without street view coverage; sampling points are placed every 250 m along the road network; Coverage Ratio = Covered points / All points.* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

between views and the choice of generative framework. Early studies demonstrated that GAN-based models are effective for viewpoint transformation. Regmi and Borji (2018), for example, showed that conditional GANs can synthesize plausible street view panoramas from satellite patches, although limited structural priors sometimes resulted in distortions. Later geometry-aware approaches strengthened this mapping. Methods incorporating height-based projection or differentiable projection (Lu et al., 2020; Shi et al., 2022) improved alignment and reduced artifacts while often remaining within the GAN family, indicating that geometry modeling and generative mechanisms are complementary rather than competing.

As geometric reasoning deepened, researchers introduced 3D and volumetric representations to address occlusion and the separation between satellite and ground perspectives. Sat2Vid (Li et al., 2021) and Sat2Density (Qian et al., 2023) leveraged point clouds, density fields, and radiance volumes to enhance structural consistency, although these approaches still faced challenges in capturing fine textures and region-specific appearance.

More recently, diffusion-based models have advanced perceptual realism through richer conditioning mechanisms. Models such as CrossViewDiff (Li et al., 2024) and Satellite to GroundScape (Xu and Qin, 2025) integrate cross-view attention, structural priors, and location-aware cues to better capture geographic identity, yet systematic generalization across diverse climates and unseen cities remains limited. Subsequent frameworks have improved control and scalability but continue to struggle with fine-grained stylistic fidelity.

Overall, progress in this field reflects a gradual improvement from basic viewpoint translation to more sophisticated geometric and stylistic reasoning. GANs remain effective for reliable perspective conversion, while diffusion models offer stronger capabilities for incorporating semantic and geographic identity information. This complementarity motivates a two-step framework in which a GAN handles viewpoint transformation and a diffusion model refines appearance to preserve region-specific geographic identity.

2.2. Image identity transfer and control

In parallel, the literature on image identity transfer provides valuable principles for controlling image appearance while maintaining semantic or geometric structure. The seminal work of Gatys et al. (2016) introduced neural style transfer using Gram matrices of convolutional features, demonstrating how content and identity could be disentangled. Building on this, translation frameworks such as Pix2Pix (Isola et al., 2017) and CycleGAN (Zhu et al., 2017) generalized identity transfer to supervised and unsupervised domain adaptation, making it possible to translate between different environments or visual domains while keeping core content intact.

These ideas have inspired the design of controllable generative models. Recent works leverage perceptual losses, semantic maps, and auxiliary structural inputs to enforce fidelity while adapting identity. For example, autoregressive and diffusion-based pipelines, such as Streetscapes (Deng et al., 2024), demonstrate that urban appearance can be flexibly modified through text prompts when structural cues (maps, heights, segmentation) are provided. Similarly, ControlNet-like frameworks (Zhao et al., 2023) have proven that composable conditioning with edges, depth, or segmentation maps enables fine-grained control over generated outputs, while recent vision–language models such as InstructBLIP (Li et al., 2023) provide high-quality semantic descriptions that further strengthen text-guided controllability.

Despite these advances, most identity transfer and controllable synthesis methods address aesthetic or environmental attributes such as lighting, weather, or seasonality, rather than region-specific geographic identity. For instance, works demonstrate the ability to “swap” urban appearances (e.g., Paris layout with New York identity), but they rarely quantify fidelity to real-world materials, vegetation, or signage tied to specific geographies. Moreover, systematic evaluation of identity preservation and generalization across held-out cities or climates remains limited. This gap is particularly important in the context of SVI generation, where realistic geographic identity is essential for downstream urban analysis tasks, especially in underrepresented regions.

3. Methodology

We introduce the GeoIdentity-Sat2Street framework designed for generating street view imagery from satellite view imagery data, with the main objective of achieving realistic synthesis while preserving geographic identity. The proposed framework is organized into two steps: the first step is a viewpoint transformer that converts satellite view imagery into the street view perspective, and the second step is the Geographic Identity and Texture-Guided Generator as illustrated in Fig. 2. The details of each step are described below.

3.1. Satellite view imagery to street-view perspective converter

To address satellite-view to street view perspective conversion, we develop a perspective conversion module. It first applies a polar transformation to the satellite input, mapping the top-view imagery into an intermediate representation that approximates a street view perspective. Based on this transformed representation, a generative adversarial framework is employed to synthesize street view panoramas.

3.1.1. Polar transformation

A central difficulty in cross-view generation lies in the drastic difference in coordinate systems: satellite images are captured in Cartesian coordinates (x,y), while street views follow a streetview perspective. To reduce this disparity, we apply a polar coordinate transformation to the satellite imagery. This idea follows earlier studies on cross-view matching, which demonstrated that polar reparameterization can effectively align satellite and street view domains by transforming circular structures into horizontal lines and radial structures into vertical lines (Shi et al., 2020).

Formally, given a satellite image of size $W_s \times H_s$, the pixel coordinates $\left(x_i^s, y_i^s\right)$ in the original image and $\left(x_i^{ps}, y_i^{ps}\right)$ in the polar-transformed image of size $W_{ps} \times H_{ps}$ are related by:

$$x_s^i = \frac{W_s}{2} + \frac{W_s}{2}\sin\left(\frac{2\pi}{W_{ps}}x_i^{ps}\right), y_s^i = \frac{H_s}{2} - \frac{H_s}{2}\cos\left(\frac{2\pi}{W_{ps}}x_i^{ps}\right) \quad (1)$$

In this formulation, circular structures in the satellite view (e.g., roundabouts) are mapped into horizontal lines, while radial structures become vertical lines. For instance, the north-line through the image center corresponds to the vertical midline in the polar-transformed representation. As shown in Fig. 3, this transformation converts a 512 × 512 square satellite image into a 512 × 1024 panoramic-like image that already resembles a street view layout.

Although this preprocessing step aligns the overall arrangement of objects, it cannot perfectly recover missing elements. Therefore, an additional generative step is necessary to synthesize a realistic street view.

3.1.2. Generative adversarial networks

To further narrow the gap between polar-transformed images and realistic street views, we adopt a conditional Generative Adversarial Network (GAN). The generator G takes the polar-transformed input I_{ps} and outputs a synthesized street view panorama $\widehat{I}_g = G(I_{ps})$. The discriminator D evaluates both generated images $\widehat{I}_g$ and real ground-truth images I_g, providing adversarial feedback that drives the generator to produce photo-realistic results.

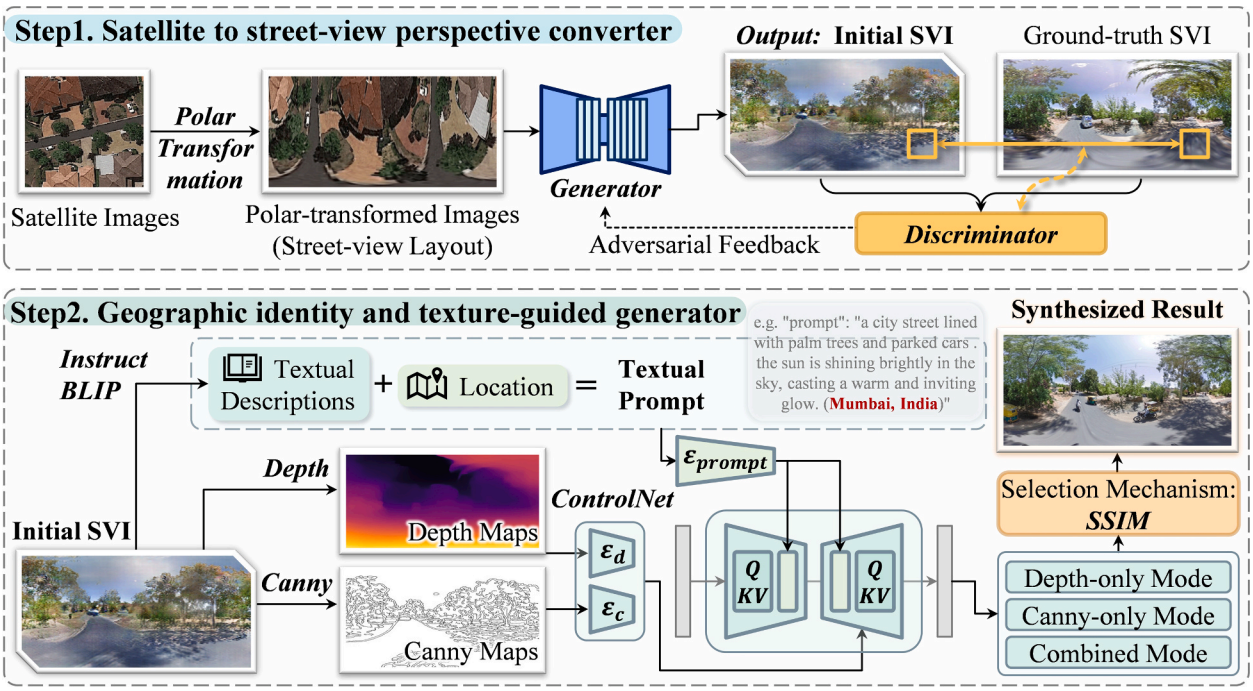

Fig. 2. The framework for geographic identity preserving street view imagery generation. *Note: Step 1 converts satellite images into an Initial SVI that provides the basic street view geometry. Step 2 refines this Initial SVI using depth, edges, text descriptions and location cues to generate a street view imagery that preserves geographic identity.*

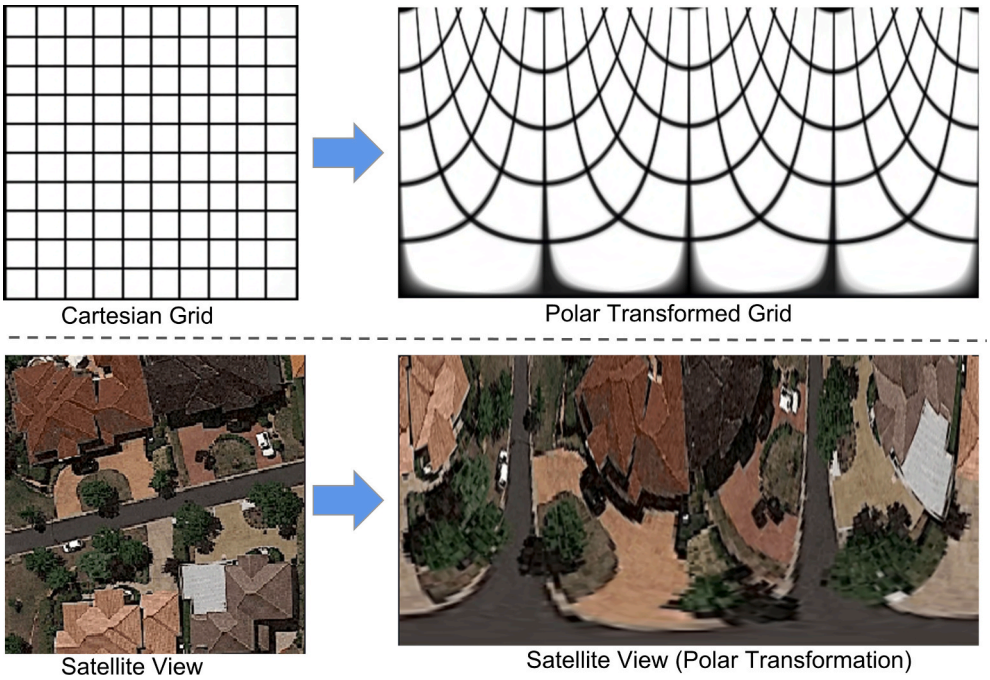

Fig. 3. A schematic of the polar transformation. *Note: The upper section shows the conversion from Cartesian to polar grids, and the lower section illustrates how satellite imagery is reparameterized into a panoramic-like layout resembling street view perspective.*

Generator Architecture. Our generator G adopts a U-Net (Ronneberger et al., 2015) backbone with residual blocks (He et al., 2016). The encoder consists of three residual downsampling blocks, each reducing spatial resolution by a factor of four, followed by six residual blocks in the bottleneck for latent feature refinement. The decoder contains three residual upsampling blocks to reconstruct the output at the same resolution as the input polar image I_{ps}. Skip connections between encoder and decoder layers are employed to facilitate gradient flow and preserve spatial details. Instance normalization (Ulyanov et al., 2017) is applied after each residual block, while spectral normalization (Miyato et al., 2018) is used after each convolutional layer to stabilize training.

Discriminator Architecture. The discriminator D follows a PatchGAN design (Isola et al., 2017), which classifies overlapping image patches as real or fake rather than evaluating the entire image globally. This patch-based strategy is particularly well-suited for street scenes, where repeated patterns such as road textures, building facades, and vegetation dominate local image statistics. By focusing on fine-grained

details instead of global coherence, the discriminator provides strong supervision for generating realistic local structures.

3.2. Geographic identity and texture-guided generator

After obtaining the initial street view panoramas from the previous step, we further refine the results to ensure geographic fidelity and textural realism. To achieve this, we introduce a geographic identity and texture-guided generator, which integrates semantic-textual guidance, location metadata, and structural priors into a diffusion-based generation framework.

During training, the diffusion model learns from large-scale street view imagery that serves as reference samples for learning realistic textures and semantic patterns. These training images are used to derive captions, location tags, and structural priors that guide the diffusion process. During inference, however, the model does not retrieve reference images for the target location. Instead, conditioning signals are constructed from the generated street view imagery produced by the previous step.

3.2.1. Text-design with BLIP and location

To ensure conditioning inputs are semantically aligned with generated images, we design a text pipeline that integrates visual descriptions with geographic metadata. During training, captions are generated from reference street view imagery in the dataset to provide semantic supervision for the diffusion model. For each image I, we use InstructBLIP (Li et al., 2023) to produce detailed captions covering road types, lane counts, buildings, vehicles, pedestrians, vegetation, weather, and lighting. These descriptions are then compressed into concise sentences with LaMini-T5 (Wu et al., 2023) to reduce redundancy and improve fluency.

In addition to semantic content, we incorporate location metadata to provide contextual geographic cues. Specifically, a filename-to-city mapping is constructed based on the dataset hierarchy, and each caption is appended with its corresponding location in the format (City, Country). When no mapping is available, the tag "unknown" is used as a placeholder. Formally, the final textual representation for an image is defined as:

$$\text{Caption}(I) = C(B(I)) + \ell_{c,k} \tag{2}$$

where $B(I)$ denotes the description generated by InstructBLIP, $C(\cdot)$ represents the LaMini-T5 compression function, and $\ell_{c,k}$ is the location tag corresponding to city c and country k. This formulation ensures that the textual representation captures both visual semantics and geographic context, which provides richer conditions for downstream generation.

During inference, the same captioning pipeline is applied to the intermediate street view imagery generated from the perspective converter, producing textual prompts that condition the diffusion model.

3.2.2. Texture-guided generation with ControlNet

While text and location prompts encode semantic and stylistic cues, they do not explicitly constrain the structural layout of generated images. To address this limitation, we adopt ControlNet (Zhang and Agrawala, 2023), a neural architecture designed to enhance diffusion models with explicit structural priors. By conditioning on edge maps, depth maps, or other geometric annotations, ControlNet ensures that generated images adhere to desired spatial constraints while maintaining semantic fidelity. During training, structural maps such as Canny edge maps and depth maps are extracted from the reference street view imagery to guide the diffusion model in learning spatial structure. During inference, these structural maps are computed from the intermediate street view imagery generated in the perspective converter.

In our framework, we use ControlNet to integrate Canny edge maps and depth maps into the generation process. For a given input image I, the structural maps are extracted as:

$$u_c = C(I), u_d = D(I) \tag{3}$$

where $C(\cdot)$ denotes the Canny edge detection function and $D(\cdot)$ denotes the depth estimation function. Conditioned on both the textual prompt p and the structural maps $\{u_c, u_d\}$, the ControlNet-augmented generator produces three candidate outputs:

$$\widehat{I}^{(c)} = G_{\text{ctrl}}(p; u_c), \widehat{I}^{(d)} = G_{\text{ctrl}}(p; u_d), \widehat{I}^{(b)} = G_{\text{ctrl}}(p; \{u_c, u_d\}) \tag{4}$$

where $\widehat{I}^{(c)}$ corresponds to the Canny-only mode that emphasizes edge sharpness and local details, $\widehat{I}^{(d)}$ corresponds to the Depth-only mode that enforces global geometry and spatial coherence, and $\widehat{I}^{(b)}$ is the Combined mode that integrates both cues for balanced structural sharpness and geometric realism.

Although ControlNet improves semantic-structural alignment, the quality of generated images can still vary across modes. To further ensure robustness, we introduce a structural similarity (SSIM)-based selection mechanism (Wang et al., 2004). For each candidate, we compute its SSIM score with respect to the input:

$$s_S = \text{SSIM}\left(\widehat{I}^{(S)}, I\right), S \in \{c, d, b\} \tag{5}$$

where s_S denotes the SSIM score of the candidate image $\widehat{I}^{(S)}$ relative to the input I. The final output is selected as:

$$S^* = \arg\max_{S \in \{c,d,b\}} s_S, \widetilde{I} = \widehat{I}^{(S^*)} \tag{6}$$

where S^* is the mode that achieves the highest SSIM score, and $\widetilde{I}$ is the retained image. This strategy allows the model to automatically retain the candidate that best preserves the structural fidelity of the input while achieving high perceptual quality.

4. Experiment

To comprehensively evaluate the effectiveness and generalization ability of our framework, we conduct experiments from three aspects: generation quality, geographic identity preservation, and a GPT-based evaluation protocol. We compare our method with InstructPix2Pix, CrossMLP, and ComingDownToEarth on CVUSA, CVACT, MultiCities Dataset, with details provided below.

4.1. Dataset details

To evaluate the effectiveness and generalization ability of our method, we conduct experiments on three types of datasets: two widely used cross-view benchmarks, a newly collected multi-city dataset for geographic identity preservation, and a large-scale global dataset for downstream evaluation.

CVUSA and CVACT. We adopt two standard cross-view benchmarks, CVUSA (Workman et al., 2015) and CVACT (Liu and Li, 2019). CVUSA contains 44,416 satellite–street view pairs from diverse U.S. regions, with 35,532 for training and 8,884 for validation. Street panoramas are provided at 1024 × 512 resolution, while satellite images are cropped to 750 × 750 and rescaled to 256 × 256. CVACT extends CVUSA to 137,218 pairs collected in Canberra, with 34,416 for training, 10,000 for validation, and 92,802 for testing. Its panoramas (1024 × 512) are paired with 512 × 512 satellite images, and precise UTM coordinates allow fine-grained evaluation. A retrieval is considered correct if the satellite location is within 5 m of ground truth, acknowledging that multiple satellite views may map to one street view imagery.

MultiCities Dataset. To assess geographic identity preservation, we construct MultiCities Dataset, a benchmark of 50,000 paired satellite–street view images from five representative cities across five continents: Mumbai (India, Asia), Amsterdam (Netherlands, Europe), New

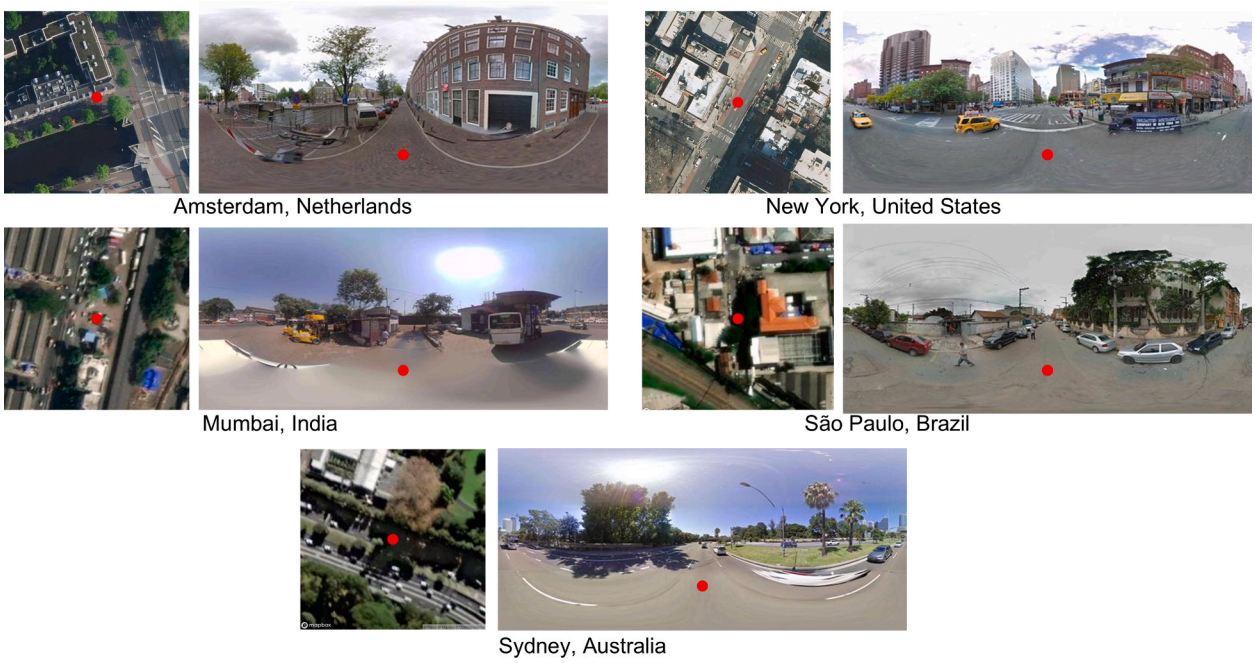

Fig. 4. Examples from the MultiCities Dataset across five cities. *Note: Each pair shows a satellite image (left) and the corresponding street view panorama (right) from Amsterdam, New York, Mumbai, São Paulo, and Sydney. Red dots indicate correspondence points between satellite and street view imagery.* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

York (United States, North America), São Paulo (Brazil, South America), and Sydney (Australia, Oceania). Each city contributes 10,000 pairs, with a 70%/30% split into training and testing sets. The street view panoramas are provided at a resolution of 1024 × 512, while the corresponding satellite images are cropped and rescaled to 512 × 512. This design ensures that the dataset captures both global diversity and local structural characteristics, enabling systematic evaluation of geographic identity preservation (sampling examples are shown in Fig. 4).

Following the data acquisition framework of the BuildingView series (Li et al., 2026b; Li et al., 2024), the dataset is constructed through a two-stage automated data acquisition pipeline. First, OpenStreetMap (OSM) road network data are downloaded within predefined city bounding boxes. For large geographic regions, the system automatically partitions the area into multiple subregions for parallel processing and merges the results afterward to ensure data completeness and consistency. The road geometries are then projected into a local UTM coordinate system, and candidate sampling points are generated by interpolating along each road LineString at fixed spatial intervals (250 m). Spatial filtering and minimum-distance constraints are applied to avoid redundant or overly dense samples, ensuring that sampling points are uniformly distributed across the road network. For each sampled location, the Google Street View API is queried to retrieve the nearest panorama within a predefined radius. The panorama tiles are downloaded and stitched into full-resolution street-view images, while metadata such as panorama ID, capture time, distance to the sampling point, and storage path are recorded in structured JSONL files. Correspondingly, satellite imagery is retrieved using the Mapbox Static API at the same geographic coordinates and resolution, ensuring precise one-to-one alignment between satellite and street view imagery.

The MultiCities Dataset captures substantial morphological diversity across sampled locations, reflecting variations in road density and road hierarchy across different urban environments. As illustrated in Fig. 5, the sampled locations span a wide range of road density conditions, covering dense urban centers, mixed-use urban areas, and lower-density suburban environments. In addition, the road hierarchy distribution derived from OpenStreetMap classifications shows that the dataset includes a diverse set of road types, including residential streets, service/access roads, minor roads, pedestrian and cycling paths, major roads, and highways. Together, these characteristics demonstrate that the dataset captures diverse streetscape environments and urban morphologies, supporting the training and evaluation of cross-view generation models under varied urban conditions.

4.2. Implementation details

All experiments are conducted on a single NVIDIA GPU with ≥24 GB VRAM, with mixed-precision training enabled. Our implementation is based on PyTorch 2.x, HuggingFace Transformers 4.x, and Diffusers 0.2x libraries. For text processing, we employ InstructBLIP (flan-t5-xl) to generate detailed scene descriptions and LaMini-T5-738 M to compress them into single sentences. For data acquisition, street view panoramas are downloaded directly from Google Street View and stitched into full-resolution images, while the corresponding satellite images are retrieved from the Mapbox API at a resolution of 512 × 512. Each image pair is stored with filenames encoding latitude, longitude, and a unique panoid identifier to ensure reproducibility.

4.3. Baseline and evaluation metrics

To assess the generalization ability of our proposed framework, we evaluate against representative baselines and employ a comprehensive suite of metrics. These metrics span pixel-level similarity, structural and perceptual fidelity, and geographic identity preservation, complemented by a GPT-based evaluation protocol that aligns with human judgment. Together, they provide a multi-perspective understanding of model performance in cross-view image generation.

4.3.1. Baseline

We compare our method with a traditional image-to-image editing model, InstructPix2Pix (Brooks et al., 2023), and two specialized cross-view synthesis baselines, CrossMLP (Ren et al., 2021) and Coming Down to Earth (Toker et al., 2021). The detailed comparison of these baselines

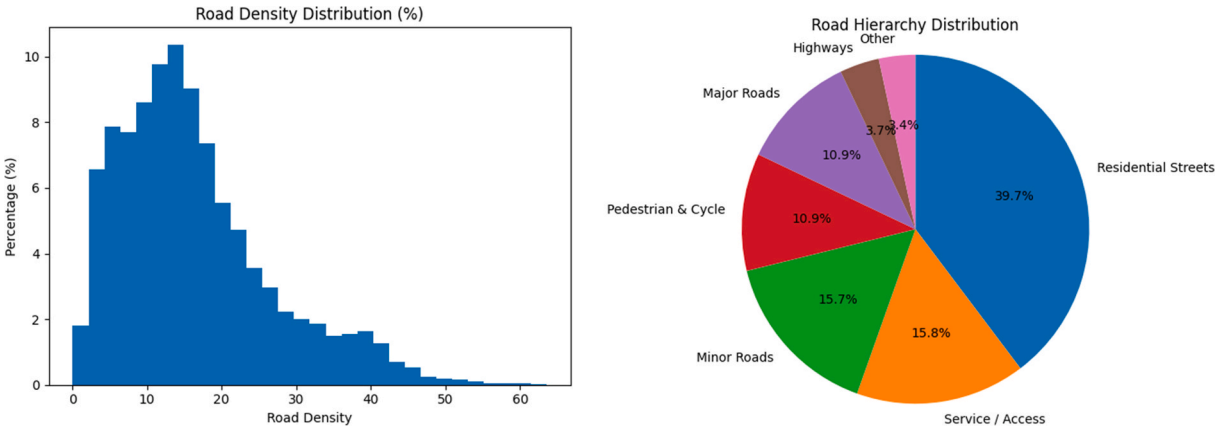

Fig. 5. Morphological diversity of the MultiCities Dataset across sampled road environments.

Table 1
Comparison of baseline methods.

Baseline	Task / Domain	Model Architecture
InstructPix2Pix (Brooks et al., 2023)	Image editing via natural language instructions	Conditional latent diffusion model with dual conditioning (input image + instruction text)
CrossMLP (Ren et al., 2021)	Cross-view image translation (satellite-street view)	Replace convolutional layers with spatial MLPs to model global spatial relationships between views
Coming Down to Earth (Toker et al., 2021)	Cross-view geo-localization & image synthesis	Jointly generate street view images from satellite imagery while optimizing retrieval alignment

in terms of task domain and model architecture is summarized in Table 1. All baseline models are implemented using their official codebases to ensure a fair comparison. For the Coming Down to Earth model, we follow the original training protocol and trained the model on the CVUSA and CVACT datasets using a batch size of 4 and an input resolution of 1024×512. The generator adopts a local architecture with 64 feature channels, 3 downsampling layers, and 6 residual blocks. The learning rate is set to 1×10^{-5}. For CrossMLP, which requires semantic segmentation masks as additional inputs, we generate the masks using a DeepLabv3+ model pretrained on the Cityscapes dataset. All preprocessing and training settings follow the configurations provided in the original implementations whenever applicable.

4.3.2. Generation quality evaluation metrics

To evaluate the generalization quality of our method, we adopt a set of widely used metrics at three complementary levels.

At the low level, we employ Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Multi-Scale SSIM (MS-SSIM). The PSNR is defined as:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{7}$$

where MAX_I is the maximum possible pixel value and MSE is the mean squared error between the generated and ground-truth images. SSIM measures structural similarity:

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{8}$$

where μ_x, μ_y are mean intensities, σ_x^2, σ_y^2 are variances, and σ_{xy} is the covariance.

MS-SSIM extends SSIM across multiple scales by progressively downsampling the images and combining luminance, contrast, and structure components.

At the edge level, we use Intersection-over-Union (IoU) and its variants. Edge-IoU is computed on extracted edges, while mIoU evaluates semantic segmentation overlaps:

$$IoU = \frac{|P \cap G|}{|P \cup G|} \tag{9}$$

where P is the predicted region (or edges) and G is the ground truth. The mean IoU averages this score across all classes or edge maps.

At the perceptual level, we employ Learned Perceptual Image Patch Similarity (LPIPS) and Kernel Inception Distance (KID). LPIPS compares deep feature embeddings:

$$LPIPS(x,y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left| w_l \odot \left(\widehat{f^l_{x,h,w}} - \widehat{f^l_{y,h,w}} \right) \right|_2^2 \tag{10}$$

where f^l are normalized deep features from a pretrained network and w_l are learned weights. KID measures distributional similarity between generated and real image features:

$$KID = \frac{1}{n(n-1)} \sum_{i \neq j} k\left(f(x_i), f(x_j)\right) + \frac{1}{m(m-1)} \sum_{i \neq j} k\left(f(y_i), f\left(y_j\right)\right) - \frac{2}{nm} \sum_{i,j} k\left(f(x_i), f\left(y_j\right)\right) \tag{11}$$

where $f(\cdot)$ denotes Inception features and k is a polynomial kernel.

These complementary metrics enable a comprehensive evaluation across pixel accuracy, structural alignment, and perceptual realism.

4.3.3. Geographic identity preservation evaluation metrics

To evaluate the preservation of geographic identity across cities, we design three complementary metrics. The average inter-city identity distance is computed as

$$D_{\text{avg}} = \frac{1}{|C|(|C|-1)} \sum_{i \neq j} |f_i - f_j|_2 \tag{12}$$

where C denotes the set of cities and f_i is the mean identity feature vector of city i. Larger values indicate stronger identity separability across cities. The variance of identity distances is given by

$$\text{Var}(D) = \frac{1}{M} \sum_{i \neq j} \left(|f_i - f_j|_2 - D_{\text{avg}} \right)^2 \tag{13}$$

where M is the number of city pairs. Smaller values suggest that identity differences are more evenly distributed across cities. Finally, the Silhouette Score jointly evaluates intra-city compactness and inter-city separability:

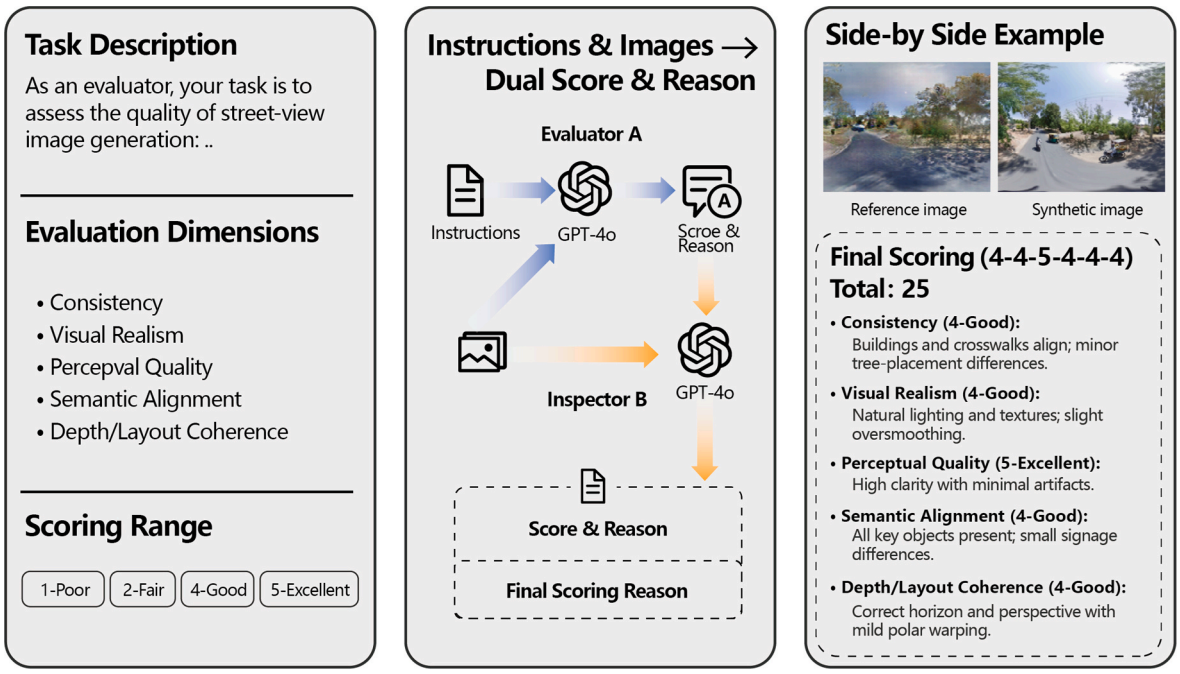

Fig. 6. Automated evaluation pipeline for street view imagery generation using GPT-4o. *Note: The framework employs task-specific meta-prompts and evaluation criteria to guide GPT-4o as Evaluator A, which generates initial scores and reasoning based on input instructions and image samples. These results are then re-assessed by GPT-4o as Inspector B, who verifies scoring appropriateness and determines the final evaluation outcome.*

$$S = \frac{1}{N}\sum_{k=1}^{N}\frac{b(k)-a(k)}{\max\{a(k),b(k)\}} \tag{14}$$

where $a(k)$ is the average distance of sample k to all other samples within the same city, and $b(k)$ is the minimum average distance of k to samples in any other city. Higher values reflect clearer clustering structure and stronger preservation of geographic identity.

To further examine whether the generated images preserve the semantic composition of urban scenes, we additionally introduce a semantic category similarity metric based on the distribution of segmentation categories. Specifically, we first obtain semantic segmentation maps for both generated and ground-truth images using a pre-trained Cityscapes segmentation model (Chen et al., 2018), which predicts the standard 19 semantic classes and achieves a mean Intersection-over-Union (mIoU) of 0.879. For interpretability, these 19 classes are further aggregated into seven broader urban semantic groups, including transport surface (road and sidewalk), built environment (building, wall, and fence), street furniture (pole, traffic light, and traffic sign), greenery (vegetation and terrain), sky, human (person and rider), and vehicle (car, truck, bus, train, motorcycle, and bicycle).

For each image k and semantic category c, we compute the category area ratio $r_{k,c}$, defined as the proportion of pixels belonging to category c. The semantic similarity between generated and ground-truth images is then measured using the mean absolute difference of category ratios:

$$\mathrm{MAE}_c = \frac{1}{N}\sum_{k=1}^{N}\left|r_{k,c}^{\mathrm{gen}} - r_{k,c}^{\mathrm{gt}}\right| \tag{15}$$

where $r_{k,c}^{\mathrm{gen}}$ and $r_{k,c}^{\mathrm{gt}}$ denote the area ratios of category c in the generated and ground-truth images, respectively, and N is the total number of image pairs. To facilitate interpretation, we further define a semantic similarity score as

$$\mathrm{Sim}_c = 1 - \mathrm{MAE}_c \tag{16}$$

where higher values indicate stronger consistency between generated and real images in terms of semantic category composition.

4.3.4. GPT-based evaluation

We adopt a GPT-based evaluation framework (Li et al., 2024) that uses LLMs as automatic evaluators (Fig. 6), offering a scalable and reproducible alternative to human annotation while remaining aligned with human judgment. The framework employs chain-of-thought prompting, in-context examples, and multiple evaluators to ensure reliable, unbiased scoring.

Specifically, we evaluate five key dimensions: Consistency, Visual Realism, Perceptual Quality, Semantic Alignment, and Depth/Layout Coherence, each scored on a 5-point Likert scale from 1 (poor) to 5 (excellent). Consistency measures whether the generated content respects the structural layout of street view imagery, including roads, landmarks, and building organization. Visual Realism assesses the plausibility of the visual appearance such as color, texture, lighting, and shading. Perceptual Quality captures overall perceptual fidelity, including sharpness, clarity, and visual comfort. Semantic Alignment evaluates whether semantic elements (e.g., cars, trees, buildings) align with expected scene semantics. Depth/Layout Coherence examines geometric and spatial structure, ensuring that depth relationships and global arrangements are coherent.

5. Result and analysis

To comprehensively assess the effectiveness of our proposed framework, we conduct evaluations from four perspectives: generalization quality, preservation of geographic identity, GPT-based human-aligned assessment, and ablation studies, as detailed in the following subsections.

5.1. Evaluating generalization quality

To evaluate the generalization capability of our framework, we conduct experiments on two widely used cross-view benchmarks, CVUSA and CVACT, which provide diverse satellite–street view pairs.

Table 2
Quantitative results on the CVUSA benchmark for evaluating generalization quality.

CVUSA	Metric	Ours	Comingdowntoearth	InstructPix2Pix	Crossmlp	Best
Low level	PSNR	14.565	14.032	12.501	**15.119**	Crossmlp
Low level	SSIM	**0.427**	0.233	0.339	0.356	**ours**
Low level	MS-SSIM	**0.438**	0.363	0.236	0.364	**ours**
Edge level	Edge_IoU	0.131	**0.153**	0.056	0.077	comingdowntoearth
Edge level	mIoU	0.052	**0.054**	0.040	0.055	comingdowntoearth
Perceptual Level	LPIPS	**0.345**	0.355	0.539	0.442	**ours**
Perceptual Level	KID	**0.086**	0.066	0.040	0.055	**ours**

Table 3
Quantitative results on the CVACT benchmark for evaluating generalization quality.

CVACT	Metric	Ours	Comingdowntoearth	InstructPix2Pix	Crossmlp	Best
Low level	PSNR	14.908	14.296	13.303	**15.575**	Crossmlp
Low level	SSIM	**0.537**	0.385	0.438	0.479	**ours**
Low level	MS-SSIM	**0.514**	0.410	0.332	0.509	**ours**
Edge level	Edge_IoU	**0.145**	0.102	0.069	0.043	**ours**
Edge level	mIoU	0.047	0.044	0.039	**0.051**	Crossmlp
Perceptual Level	LPIPS	**0.317**	0.360	0.457	0.546	**ours**
Perceptual Level	KID	0.059	**0.078**	0.025	0.411	comingdowntoearth

These datasets allow us to rigorously assess image quality from multiple perspectives, including pixel-level similarity, structural consistency, and perceptual realism. The evaluation is carried out using traditional metrics at three levels: low-level (PSNR, SSIM, MS-SSIM), edge-level (Edge-IoU, mIoU), and perceptual-level (LPIPS, KID). The quantitative results are presented in Tables 2 and 3.

On CVUSA, our method achieves the best SSIM (0.427) and MS-SSIM (0.438), and the lowest LPIPS (0.345), with superior mIoU (0.054) for structural layout. On CVACT, it again leads in SSIM (0.537), MS-SSIM (0.514), and LPIPS (0.317). While CrossMLP attains the highest PSNR (15.575) and ComingDownToEarth the lowest KID (0.078), our framework consistently outperforms in pixel-level similarity, structural fidelity, and perceptual realism.

Overall, these results demonstrate that our framework consistently achieves superior pixel-level similarity through higher SSIM and MS-SSIM scores, showing its strength in preserving fine details and maintaining global coherence. It also exhibits strong structural fidelity by achieving the best mIoU performance, which indicates robustness in capturing geometric layout, while balancing flexibility in scenarios where explicit geometric modeling may dominate. At the perceptual level, our framework achieves the lowest LPIPS scores across datasets, confirming that it produces visually realistic and semantically consistent street view imagery, with competitive KID results further validating its stability in distributional alignment. Importantly, the consistent gains across both CVUSA and CVACT reveal the robustness of our approach in image generation quality.

5.2. *Evaluating for preserving geographic identity*

To further investigate whether our method is capable of preserving geographic identity, we evaluate on the MultiCities Dataset (50,000 satellite–street pairs from Mumbai, Amsterdam, New York, São Paulo, and Sydney). To assess whether models capture city-specific identities, we use three complementary metrics: average inter-city identity distance, variance of identity distances, and the Silhouette Score, jointly measuring inter-city separability and intra-city compactness. The quantitative results are presented in Table 4.

Table 4
Quantitative results on MultiCities Dataset for evaluating preservation of geographic identity.

Model	Mean Inter-City Identity Distance ↑	Identity Distance Variance	Silhouette Score ↑
Ours	**0.103**	0.031	**0.222**
Comingdowntoearth	0.062	0.024	0.131
CrossMLP	0.011	0.003	≈ 0
InstructPix2Pix	0.004	0.002	−0.049

Our method achieves the highest Silhouette Score (0.222), indicating the clearest clustering of city-specific identities. It also produces a competitive Average inter-city identity distance (0.103) and Variance of identity distances (0.031), demonstrating its ability to balance both identity separability and stability across different regions. In comparison, ComingDownToEarth achieves slightly lower variance but underperforms on inter-city separation, while CrossMLP and InstructPix2Pix fail to capture meaningful geographic distinctions, with InstructPix2Pix even producing a negative Silhouette Score (–0.049).

To provide a more intuitive perspective, we also visualize the city-level distributions of generated street view imagery using t-SNE, as shown in Fig. 7. InstructPix2Pix and CrossMLP exhibit highly overlapping clusters with indistinguishable city boundaries. ComingDownToEarth partially separates cities but shows noisy overlaps. By contrast, our method produces well-separated and compact clusters for all five cities, clearly reflecting their distinctive geographic identities. These results confirm that our framework not only generates visually realistic images but also preserves meaningful geographic diversity across different regions.

To complement the quantitative and clustering analyses, Fig. 8 presents qualitative comparisons of synthesized street view imagery across five cities. For each city, we show the ground-truth panorama alongside results from ComingDownToEarth, CrossMLP, InstructPix2Pix, and our method. InstructPix2Pix often produces distorted structures and fails to preserve geographic context; CrossMLP yields blurred, unrealistic images; and ComingDownToEarth, while sharper, loses fine-grained texture details. By contrast, our method generates panoramas with clearer buildings, vegetation, and road layouts that more closely match the ground truth.

To complement the previous analyses, we additionally evaluate how well the generated images preserve the semantic composition of urban scenes. Specifically, we measure the similarity between the semantic category distributions of synthesized and ground-truth images by computing semantic similarity score (Formula 16) between their category ratio vectors, where higher values indicate closer agreement in semantic composition.

Table 5 presents the overall comparison across different methods. Our method achieves the highest similarity in most semantic categories, including transport surface (0.877), built environment (0.876), sky

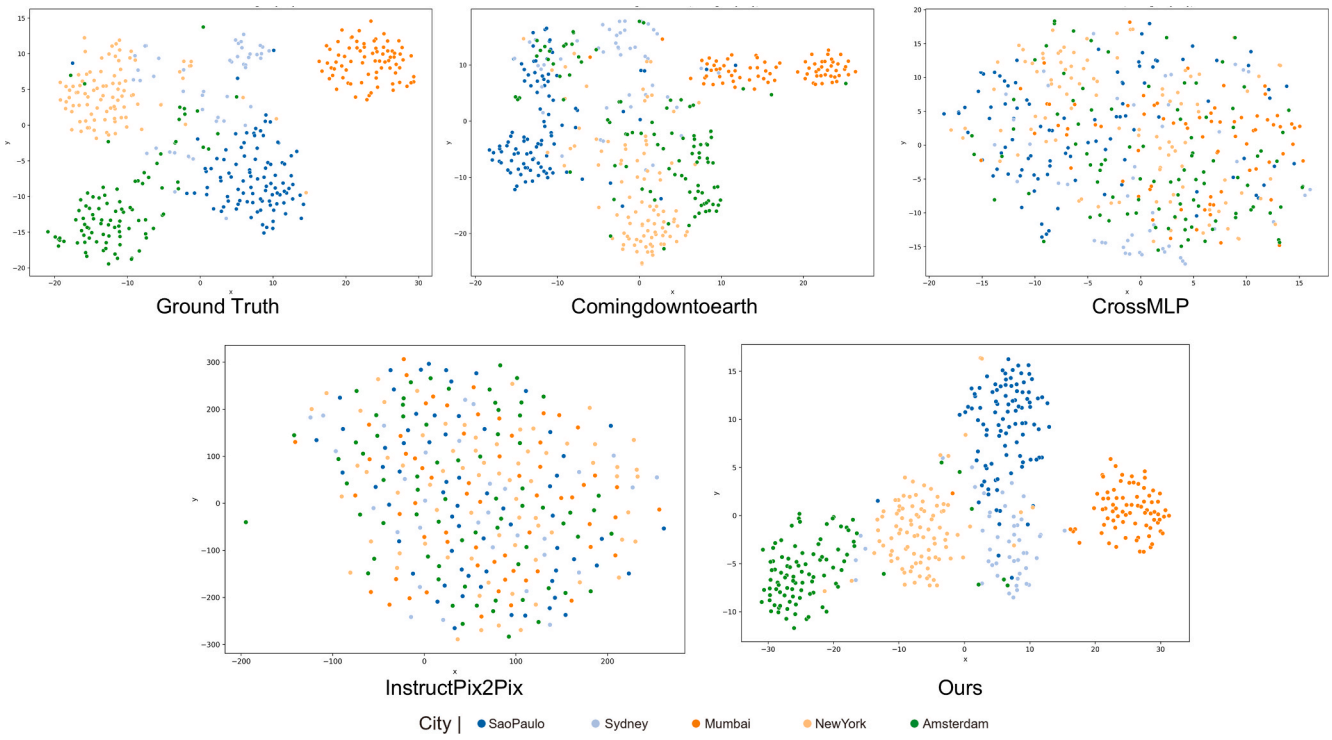

Fig. 7. T-sne visualization of generated street view imagery across five cities using different methods. *Note: Comparison of Ground truth, InstructPix2Pix, CrossMLP, ComingDownToEarth, and our method. Prior methods yield overlapping or noisy clusters, while ours produces clear, compact city-specific clusters.*

(0.909), human (0.997), and vehicle (0.847), demonstrating that the generated images faithfully preserve the structural composition of urban environments.

In addition, Table 6 reports the city-level results of ours across the five cities in the MultiCities dataset. The results show consistently high similarity scores across cities, with mean values ranging from 0.895 in Amsterdam to 0.925 in New York. This stability across geographically diverse urban environments indicates that the proposed framework not only captures city-specific visual identities but also maintains the semantic composition of key urban elements.

To further examine whether the proposed framework preserves spatial heterogeneity within cities, we conduct an additional neighborhood-level analysis based on visual embedding distances. Specifically, we construct a multi-scale evaluation framework that assesses spatial structure from both structural and textural perspectives. Using geographic coordinates, each city is partitioned into regular grids with a fixed spatial resolution (e.g., 500 m), where each grid cell represents a local neighborhood unit capturing intra-urban spatial structure. For each image, we extract visual embeddings using a pretrained CLIP model and normalize the embeddings to ensure comparability of cosine distances. Based on these embeddings, we compute two metrics: (1) the average cosine distance between images within the same grid cell (within-cell distance), and (2) the average cosine distance between images across different grid cells (between-cell distance). The ratio between these two quantities (between/within) is used as a neighborhood separability indicator. Intuitively, if a generative model excessively smooths intra-city spatial variation, the between/within ratio would decrease, indicating weaker separability between neighborhoods.

As shown in Fig. 9, the ground-truth images (GT) exhibit an average within-cell distance of 0.1969 and a between-cell distance of 0.2840, yielding a neighborhood separability ratio (between/within) of 1.4425. For the generated images (GEN), the corresponding values are 0.1565, 0.2237, and 1.4289, respectively. Although both the within-cell and between-cell distances are slightly smaller in the generated images, the separability ratio remains nearly unchanged (1.4425 vs. 1.4289), indicating that the relative spatial differentiation between neighborhoods is effectively preserved. The bootstrap distributions of the separability ratio for GT and GEN also show substantial overlap, with highly similar central tendencies and overall distribution shapes. This consistency suggests that the generative model successfully reconstructs the relative spatial separation structure among neighborhoods in the visual embedding space. In other words, the generated images maintain the spatial organization patterns observed in real urban environments, preserving fine-grained intra-city spatial heterogeneity rather than collapsing urban visual patterns.

Overall, our framework achieves the best balance between inter-city separability, intra-city compactness, and stability, as reflected in Silhouette Score and variance analysis. Clustering visualizations confirm that our method better captures city-specific identity distributions, while qualitative comparisons show sharper and more realistic panoramas. Together, these results demonstrate that our approach both preserves meaningful geographic diversity and delivers high-quality street view synthesis, setting it apart from prior methods.

5.3. GPT-based evaluation

To complement traditional and identity-preserving evaluations, we use a GPT-based framework (Li et al., 2024) to assess synthesized images across five dimensions: Consistency, Visual Realism, Perceptual Quality, Semantic Alignment, and Depth/Layout Coherence—on a five-point Likert scale, serving as a reliable proxy for human judgment. To enhance robustness and statistical reliability, we adopt a random k-fold partitioning with repeated evaluation. Image pairs are randomly divided into multiple folds, and the GPT-based scoring procedure is independently performed on each fold across several rounds. The final results report the mean scores together with their variation ranges, providing a more reliable estimate of model performance and reducing potential randomness in LLM-based evaluation.

As shown in Table 7, our method consistently achieves the strongest overall performance, with a total score ranging from 10.587 to 10.824, substantially surpassing all baselines. In particular, it achieves the highest scores in Visual Realism (2.292–2.344), Semantic Alignment

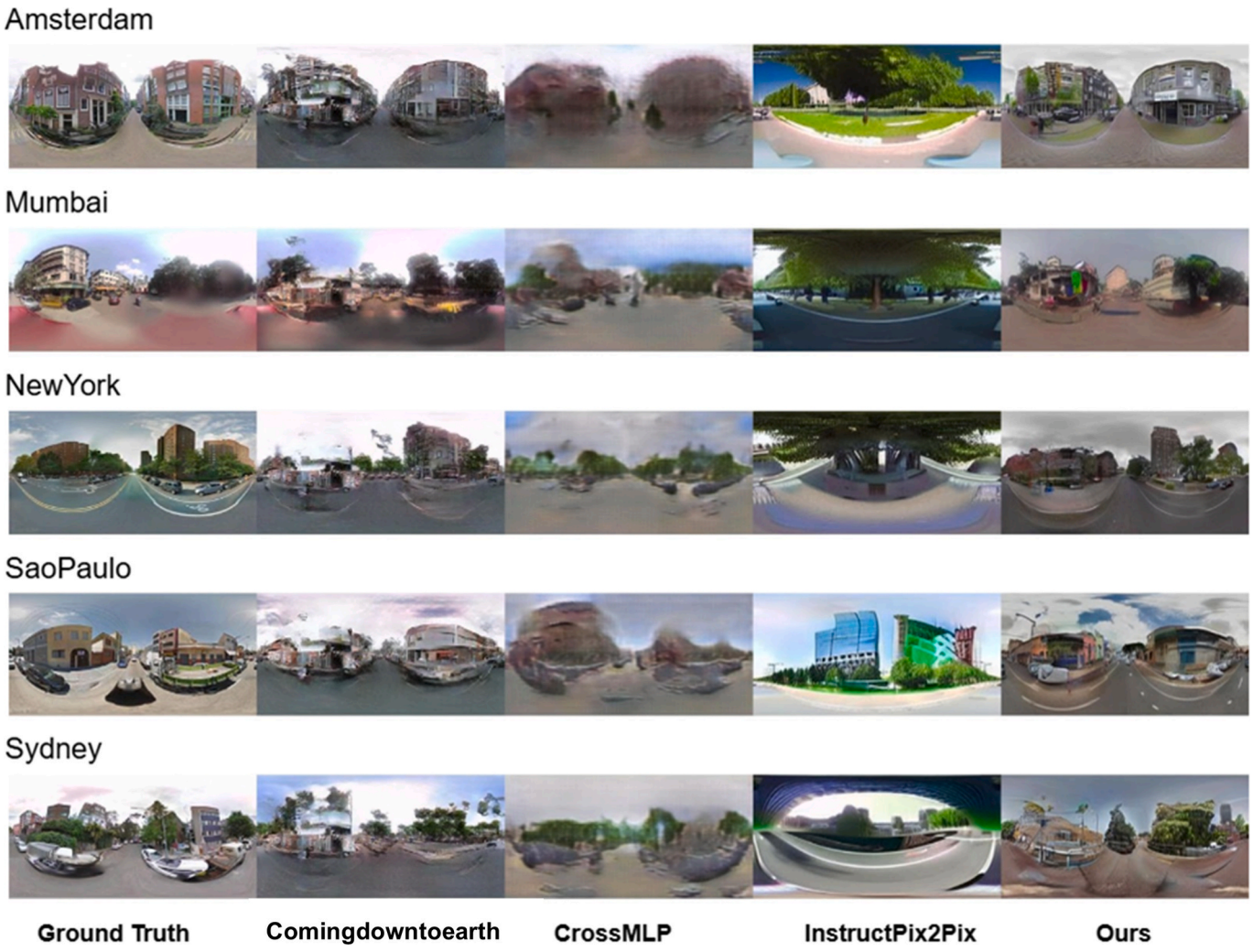

Fig. 8. Qualitative comparison of street view imagery generation across five cities in the MultiCities Dataset. *Note: Ground-truth panoramas are shown alongside results from ComingDownToEarth, CrossMLP, InstructPix2Pix, and our method for Amsterdam, Mumbai, New York, São Paulo, and Sydney.*

Table 5
Semantic composition similarity between generated and ground-truth images across semantic categories (semantic similarity score, higher is better).

Category	Crossmlp	InstructPix2Pix	Comingdowntoearth	Ours
Transport Surface	0.848	0.733	0.868	**0.877**
Built Environment	0.721	0.818	0.846	**0.876**
Street Furniture	0.996	0.986	0.995	**0.997**
Greenery	**0.898**	0.787	0.889	0.886
Sky	0.857	0.831	0.897	**0.909**
Human	0.982	0.985	0.981	**0.997**
Vehicle	0.800	0.747	0.839	**0.847**

Table 6
City-level semantic composition similarity of ours across the five cities in the MultiCities Dataset.

City	Transport Surface	Built Env	Greenery	Sky	Human	Vehicle	Mean
Amsterdam	0.841	0.860	0.893	0.900	0.996	0.778	0.895
Mumbai	0.888	0.907	0.876	0.930	0.998	0.843	0.920
New York	0.895	0.853	0.903	0.926	0.997	0.900	0.925
São Paulo	0.91	0.898	0.878	0.901	0.999	0.858	0.920
Sydney	0.832	0.855	0.866	0.878	0.996	0.863	0.898
Overall	0.877	0.876	0.886	0.909	0.997	0.847	**0.913**

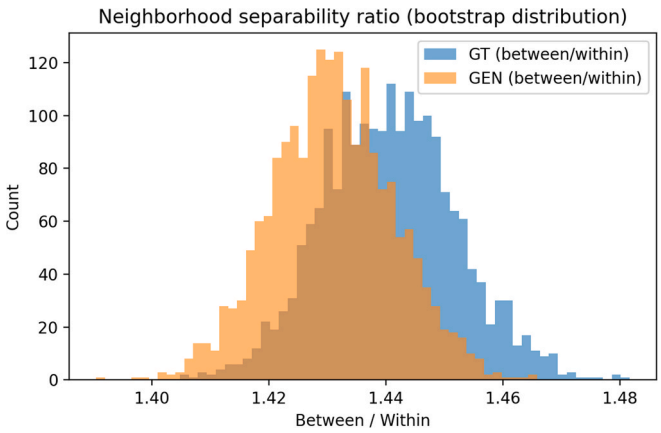

Fig. 9. Bootstrap Distribution of the Between/Within Neighborhood Separability Ratio.

(2.234–2.289), and Depth/Layout Coherence (2.379–2.439), indicating that the generated images are visually convincing, semantically accurate, and structurally coherent. Among the baselines, ComingDownToEarth demonstrates relatively stronger Consistency (1.949–2.003) but performs weaker in perceptual realism and semantic fidelity. InstructPix2Pix achieves competitive Perceptual Quality (1.326–1.382) but exhibits weaker structural coherence and semantic alignment. CrossMLP performs consistently lower across most dimensions. Overall, baselines display strengths in isolated aspects but lack balanced performance. In contrast, our method delivers consistent improvements and excels in realism, semantic fidelity, and structural coherence, demonstrating a comprehensive advantage in high-quality, geographically faithful cross-view synthesis.

5.4. Ablation study and sensitivity analysis

To better understand the contribution of different components in our framework, we conduct an ablation study by systematically removing the location text-design, ControlNet guidance, and polar transformation. The performance is evaluated across pixel-level, structural, perceptual, and geographic identity metrics, with results summarized in Table 8.

Our full model achieves the best overall balance, with the highest PSNR (12.086), SSIM (0.337), and MS-SSIM (0.228), demonstrating superior pixel-level fidelity. Removing the location text-design slightly lowers SSIM and MS-SSIM but leads to a noticeable decline in visual distinctiveness, as reflected by a negative Silhouette Score (–0.024), underscoring its role in maintaining inter-city differentiation. Excluding

Table 7
Quantitative scores of different methods under GPT-based evaluation.

Model	Consistency	Visual Realism	Perceptual Quality	Semantic Alignment	Depth/Layout Coherence	Total
Comingdowntoearth	**1.949–2.003**	1.736–1.768	1.505–1.554	2.094–2.163	1.959–2.002	9.243–9.490
CrossMLP	1.648–1.693	1.011–1.027	1.009–1.002	1.648–1.693	1.648–1.693	6.964–7.108
InstructPix2Pix	1.010–1.024	1.516–1.559	1.326–1.382	1.019–1.042	1.305–1.361	6.176–6.368
Ours	1.726–1.772	**2.292–2.344**	**1.956–1.980**	**2.234–2.289**	**2.379–2.439**	**10.587 –10.824**

Note: GPT-based evaluation shows our method achieves the highest overall score, with clear advantages in Visual Realism, Semantic Alignment, and Depth/Layout Coherence.

Table 8
Ablation study results on the contributions of location text-design, ControlNet guidance, and polar transformation.

Model	PSNR ↑	SSIM ↑	MS-SSIM ↑	Edge IoU ↑	LPIPS ↓	mIoU ↑	KID std % ↓	Silhouette ↑
Ours (A + B + C)	**12.086**	0.337	**0.228**	0.226	**0.614**	0.395	**0.450**	**0.222**
w/o Location (A + C)	11.942	**0.341**	0.223	0.246	0.618	**0.396**	0.565	–0.024
w/o ControlNet (A + B)	10.727	0.264	0.176	**0.239**	0.621	0.368	0.601	0.173
w/o Polar (B + C)	11.909	0.332	0.210	0.218	0.627	0.377	0.615	0.017

Note. "↑" indicates that a higher value corresponds to better performance, while "↓" indicates that a lower value corresponds to better performance. Module A refers to the location text-design module, B refers to the ControlNet guidance module, and C refers to the polar transformation module.

Table 9
Sensitivity analysis of training hyperparameters.

CVUSA	Metric	lr1e-05_ep10	lr1e-05_ep15	lr2e-05_ep10	lr2e-05_ep15	lr5e-05_ep10	lr5e-05_ep15
Low level	PSNR	11.8400	11.9185	11.7766	11.8210	11.6002	11.9489
Low level	SSIM	0.3354	0.3481	0.3250	0.3428	0.3235	0.335
Low level	MS-SSIM	0.2132	0.2114	0.2025	0.2020	0.1938	0.2126
Edge level	Edge_IoU	0.0912	0.0908	0.0888	0.0880	0.0921	0.0906
Edge level	mIoU	0.0451	0.0448	0.0441	0.0437	0.0425	0.045
Perceptual Level	LPIPS	0.4795	0.4802	0.4868	0.4853	0.4853	0.4808
Perceptual Level	KID	0.0499	0.0556	0.0513	0.0547	0.0537	0.0532

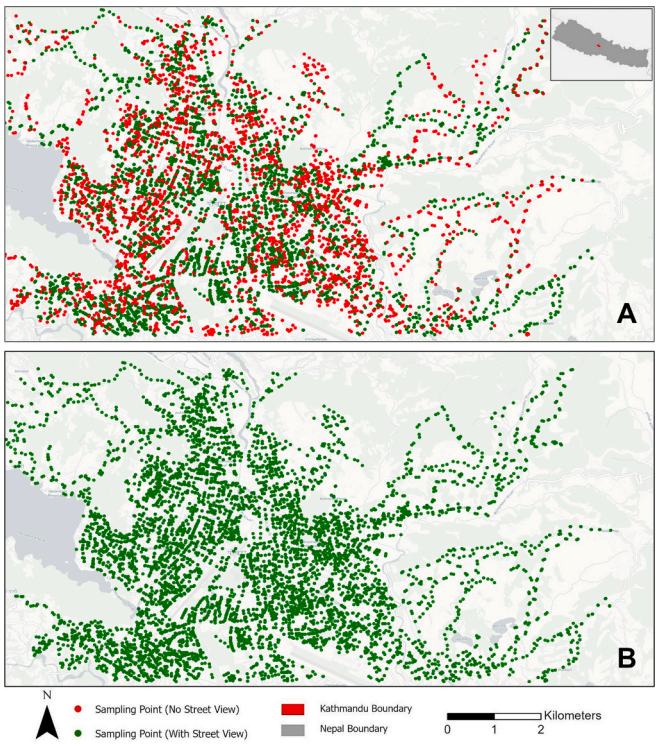

Fig. 10. Sampling points of street view coverage in Kathmandu, Nepal (before and after data completion). *Note: Among 8,917 sampled points, 71.8% (green) have imagery while 28.2% (red) remain uncovered, showing substantial gaps in Google Street View coverage. (A) Before completion; (B) After completion. Green points indicate covered areas, and red points denote uncovered areas.* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ControlNet guidance causes the most severe degradation, especially in SSIM (0.264) and LPIPS (0.621), showing its necessity for structural and perceptual fidelity. Polar transformation also weakens perceptual and identity metrics, though less severely than the other components. Taken together, these results confirm that each component contributes to complementary aspects of performance: the location text-design is indispensable for geographic identity preservation, ControlNet guidance is vital for structural and perceptual fidelity, and polar transformation further enhances visual realism. Their combined effect enables our full model to achieve a consistent and comprehensive advantage over all ablated variants.

To further examine the robustness of the training configuration, we conduct a sensitivity analysis on key hyperparameters, including the learning rate and the number of training epochs. Specifically, we evaluate several combinations of learning rates (1×10^{-5}, 2×10^{-5}, and 5×10^{-5}) and training durations (10 and 15 epochs) on a subset of 1000 samples from the CVUSA dataset to efficiently assess training stability while maintaining representative data coverage. The evaluation results are summarized in Table 9.

The results indicate that the performance variations across different hyperparameter settings are relatively limited, demonstrating that the proposed framework maintains stable performance under moderate changes in learning rate and training epochs.

6. Downstream application

To further demonstrate the practical value of our framework, we explore its application in addressing fairness and completeness issues in global street view data coverage through a case study in Kathmandu, Nepal.

As shown in Fig. 10 A, we evaluate street view coverage in Kathmandu by sampling OSM road points at 200 m intervals and retrieving Google Street View imagery within 100 m. Green points indicate coverage, whereas red points highlight substantial gaps that hinder downstream analysis. Most street view data in this region is missing, resulting in significant deficiencies in subsequent analyses. To address this issue, we use the framework trained on the MultiCities Dataset with corresponding Mapbox satellite imagery as input to synthesize realistic street view imagery that fill the missing areas in Kathmandu. As shown in Fig. 10 B, the generated results successfully complete the uncovered regions, thereby enhancing data completeness and fairness.

Furthermore, we employ t-SNE visualization to compare the distribution of generated and real street view imagery (Fig. 11). The results show that both distributions exhibit highly consistent clustering patterns, indicating that our synthesized panoramas preserve the same underlying identity characteristics as real data. This demonstrates that the framework trained on the MultiCities Dataset can effectively generalize to unseen regions, enabling its transfer to broader street view completion tasks.

To further validate the reliability of perception-based analysis derived from synthesized imagery, we additionally compare perception scores computed from real and generated street view imagery in Kathmandu. Fig. 12 shows the distribution of safety perception scores derived from both data sources. The two distributions exhibit similar patterns, suggesting that the synthesized images preserve perceptual characteristics comparable to those of real street view imagery. To statistically evaluate the similarity between the two distributions, we perform a Kolmogorov–Smirnov (KS) two-sample test. The test yields a KS statistic of 0.183 with a p-value of 0.089, indicating that there is no statistically significant evidence that the perception scores derived from generated and real images come from different distributions.

In addition, Fig. 13 presents side-by-side visual comparisons of real and synthesized street view panoramas from Kathmandu. These examples demonstrate that the generated images capture key streetscape structures, building layouts, and environmental features consistent with real scenes.

In summary, our framework not only generates consistent street view imagery but also enhances fairness by filling data gaps in underrepresented regions. Trained on the MultiCities Dataset, it generalizes well to diverse global contexts, supporting more equitable urban intelligence.

7. Discussion

7.1. Bridging the street view imagery coverage gap through identity preservation

A central insight from our study is that bridging the street view imagery coverage gap requires going beyond perspective alignment to also ensure faithful preservation of local identity. In the cross-view literature, benchmark datasets such as CVUSA (Workman et al., 2015) and CVACT (Liu and Li, 2019) have long served as standards for evaluating translation between satellite and street view imagery. However, their geographic scope is limited to a few data-rich regions. This narrow coverage overlooks the distinctive geographic identity, including architectural morphology, vegetation patterns, and urban form diversity, that we highlighted in the introduction as fundamental characteristics of global streetscapes. Recognizing and preserving these local stylistic identities is therefore essential for advancing cross-view synthesis beyond mere pixel alignment. As a result, existing work primarily advances synthesis fidelity but does not address the deeper challenge of coverage gaps in global street view data.

Our findings demonstrate that preserving geographic identity is the key to bridging both perspective and coverage gaps. Through polar transformation, our method reduces structural disparity between satellite and street views, while texture- and semantics-guided generation ensures that the synthesized images retain geographically grounded

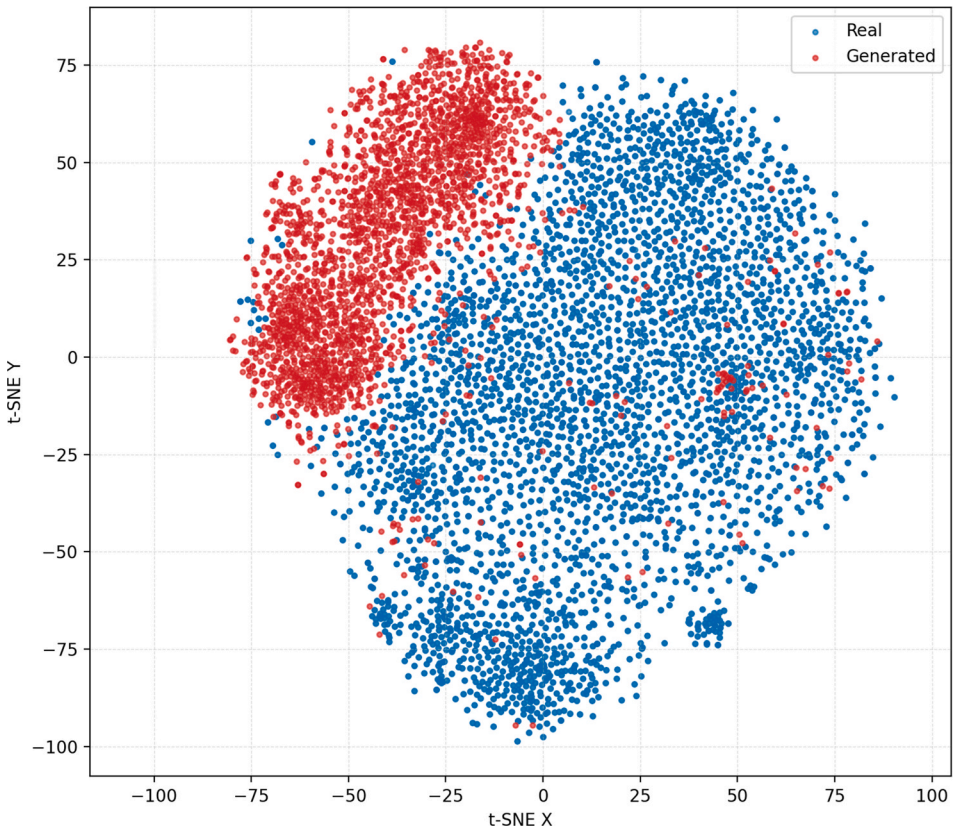

Fig. 11. T-sne visualization of real and generated street view image distributions. *Note: Blue (real) and red (generated) samples form consistent clustering patterns, showing that our synthesized panoramas capture local geographic identity comparable to real data, supporting their use in global-scale street view completion.* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

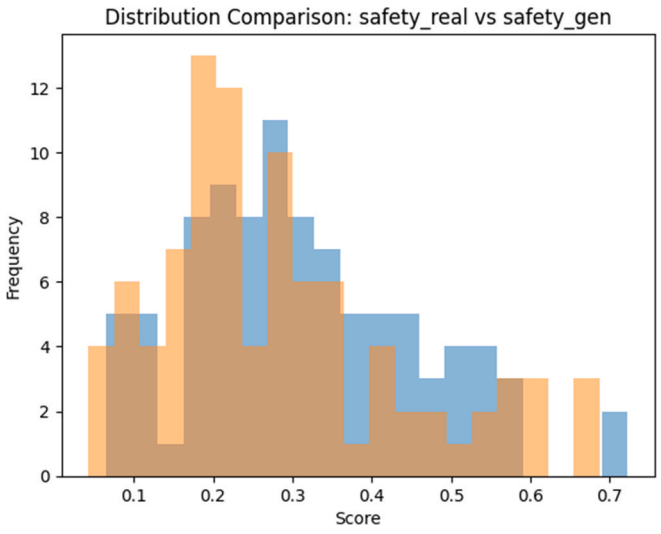

Fig. 12. Distribution comparison of perception scores derived from real and generated street view imagery in Kathmandu. *Note: The histogram shows the distribution of safety perception scores computed from real images (blue) and synthesized images (orange).* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

characteristics such as building layouts, vegetation, and streetscape textures. Moreover, the t-SNE visualization of learned embeddings (Fig. 6) reveals that scenes from geographically proximate regions cluster closely in the latent space, indicating that the model captures consistent regional identity semantics constrained by real-world geography. This pattern highlights the model's ability to internalize global urban characteristics—where nearby cities share stylistic similarities yet remain distinguishable by their local identities. Such results confirm that identity consistency under geographic constraints is not an incidental artifact, but an emergent property of our model's learned representation space.

In this way, we reframe cross-view synthesis from a problem of domain alignment toward one of data equity and representational completeness. Preserving geographic identity is not merely a visual refinement, but a mechanism for extending street view coverage in data-scarce regions, enabling more inclusive applications in safety perception, urban planning, and environmental monitoring across globally diverse environments.

7.2. *Extending Street view analytics to global south cities*

Street view–based urban analytics are shaped by deep geographic disparities. While cities in the Global North benefit from extensive coverage by platforms such as Google Street View, large portions of the Global South, particularly in Africa, South Asia, and Latin America, remain sparsely mapped or entirely excluded (Fry et al., 2020; Yin et al., 2023). This limits the visibility of diverse urban morphologies, street-level infrastructure, and environmental cues in global studies. Urban research that relies on dense visual data, including assessments of road safety, greenery, or housing quality, is therefore disproportionately skewed toward well-mapped regions.

Our framework provides a pathway to mitigate this imbalance by generating geographically consistent and identity-faithful synthetic street views in data-scarce contexts. Its high transferability allows it to generalize beyond the training distribution and effectively address coverage gaps, especially in Global South cities where data scarcity is most pronounced.

Fig. 13. Side-by-side comparison of real and synthesized street view in Kathmandu.

To further demonstrate this extensibility, we conduct a case study in Kathmandu, Nepal. Building upon the model trained on the MultiCities Dataset, we extend its application to a broader global setting using the NUS Global Streetscapes dataset (Hou et al., 2024), which contains over 10 million crowdsourced panoramic images from 688 cities across 210 countries, sourced from Mapillary and KartaView. Leveraging this large-scale dataset, our framework synthesizes realistic street view imagery that fills missing coverage in Kathmandu, thereby enhancing data completeness and fairness. We further evaluate its downstream utility through road-safety perception analysis using Place Pulse 2.0 (Zhang et al., 2021), where the supplemented imagery leads to denser and more continuous safety maps compared to the baseline with large uncovered regions (Fig. 14).

These findings demonstrate that even when trained on globally diverse datasets, our framework effectively captures local geographic identity and generalizes to unseen regions. This transferability enables its deployment for large-scale street view completion across Global South cities, promoting representational completeness, fairness, and more inclusive urban analytics worldwide.

However, it is important to clarify the applicability boundaries of generated street view imagery in downstream urban analytics. Because the proposed framework primarily preserves large-scale streetscape structure, semantic composition, and city-level visual identity, the synthesized imagery is particularly suitable for analyses that rely on macro-level visual patterns or scene-level semantics.

From the perspective of classical computer vision tasks, the generated imagery can effectively support analyses based on scene perception and semantic composition. For example, perception-based urban studies such as safety perception mapping rely primarily on overall streetscape appearance and visual context rather than precise geometric measurements (Zhang et al., 2021). Similarly, tasks such as urban greenery estimation (Zhu et al., 2023), built-environment characterization (Chen et al., 2022), and cross-city visual identity analysis mainly depend on the presence and spatial distribution of major semantic elements (e.g., vegetation, buildings, and roads), which are generally well preserved in the generated street view imagery.

In contrast, some urban analytics tasks require highly accurate geometric information extracted from street view imagery, which may not be reliably preserved in cross-view synthesized images. For instance, recent studies have used street view imagery to estimate street width and evaluate ambulance trafficability when assessing emergency medical service accessibility for older adults (Liu et al., 2025). Similarly, street view imagery has been used to derive detailed built-environment indicators such as street enclosure and greenery visibility when analyzing spatio-temporal relationships between the built environment and metro–bike travel (Wei et al., 2025). These applications typically rely on precise pixel-level geometry or the reliable detection of small-scale urban elements. Because cross-view image generation inherently involves a one-to-many mapping between satellite and street view perspectives, the synthesized imagery may not always preserve such fine-grained geometric details.

Therefore, generated street view imagery is more suitable for urban analytics tasks based on scene perception and semantic composition, while applications requiring precise geometric measurements or accurate detection of small urban objects should be applied with caution.

7.3. Limitations and future work

While our framework demonstrates strong performance in geographic identity preservation and global applicability, several limitations remain that open directions for future research.

First, although the proposed method achieves strong overall performance, it still exhibits several failure modes in fine-grained detail generation. While large-scale structural layout and city-level style are

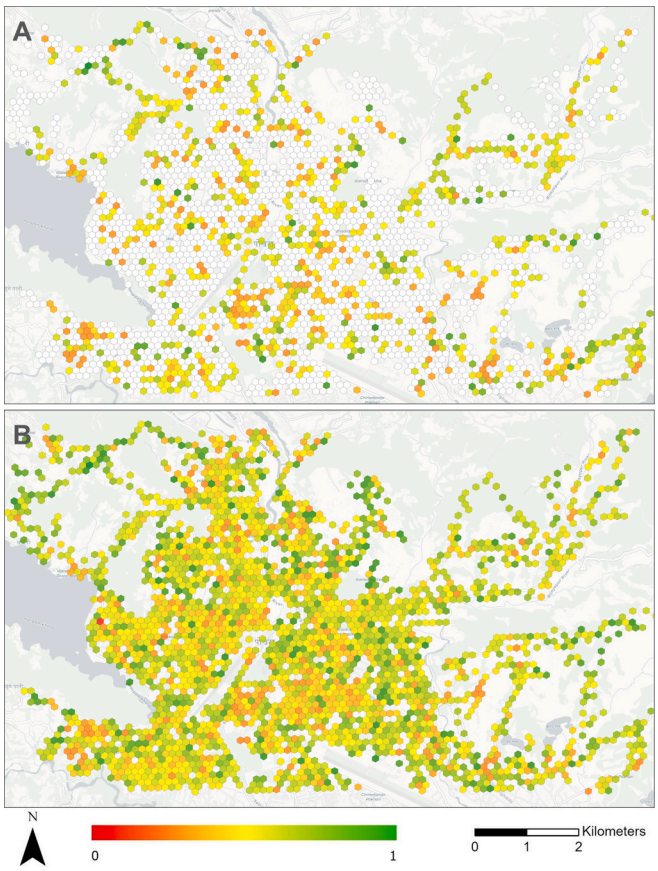

Fig. 14. Road safety perception before (A) and after (B) coverage supplementation in Kathmandu, Nepal. *Note: (A) shows sparse safety estimations with missing imagery, while (B) becomes denser and more continuous after adding generated street views.*

generally preserved, small-scale elements and local geometric consistency can occasionally degrade. As illustrated in Fig. 15, several representative failure cases are observed.

In some scenes (A), local regions exhibit severe blurring and content collapse, leading to the loss of semantic details and fine structures. In other cases (B), the model struggles to maintain geometric consistency across panoramic seams, resulting in distorted building structures and local warping artifacts. Additionally, extreme panoramic viewpoints may produce abnormal sky textures and structural distortions in open scenes (C). Local structural inconsistencies can also appear along roads and boundary objects such as fences or walls (D). These examples highlight the challenges diffusion-based models face in jointly preserving global panoramic geometry and fine-grained urban details. Such limitations mainly arise from the difficulty of modeling small-scale semantic objects while maintaining global geometric consistency in wide-field panoramic synthesis. Future work could address these issues by incorporating higher-resolution generation, stronger geometric constraints, or multimodal supervision (e.g., visual–language alignment) to improve object-level realism and structural stability.

Second, another limitation lies in the current evaluation protocol. Although the study adopts a relatively comprehensive set of metrics covering low level fidelity (PSNR, SSIM, MS-SSIM), edge level consistency (Edge IoU, mIoU), perceptual similarity (LPIPS, KID), and semantic composition similarity based on segmentation category distributions, these quantitative metrics still mainly emphasize global structure and overall scene composition. Consequently, certain fine-grained inconsistencies, such as small street objects, subtle geometric distortions, or localized structural artifacts, may remain underrepresented in the evaluation results. Future work could therefore develop more fine grained evaluation frameworks that explicitly measure object level realism and spatial consistency. For example, instance level segmentation metrics could be used to evaluate the presence and geometric accuracy of specific urban elements such as traffic signs, poles, vehicles, and street furniture. In addition, object detection based consistency measures could be applied to compare the number, spatial distribution, and relative arrangement of urban objects between generated and real images. Another promising direction is to incorporate downstream task based validation, such as accessibility estimation, streetscape perception analysis, or urban morphology classification, to evaluate whether the

Table 10
Comparison of viewpoint transformation architectures on the CVUSA dataset.

CVUSA	Metric	Conditional GAN (Ours)	SPADE-based models
Low level	PSNR	**14.5648**	13.8094
Low level	SSIM	**0.4269**	0.4192
Low level	MS-SSIM	**0.4382**	0.3288
Edge level	Edge_IoU	**0.1312**	0.1127
Edge level	mIoU	0.0515	**0.0521**
Perceptual Level	LPIPS	0.3453	**0.4238**
Perceptual Level	KID	**0.0857**	0.0784

Fig. 15. Representative failure cases of our method. *Note: Red boxes highlight representative failure regions. Typical errors include severe local blurring and content collapse (A), distorted building geometry near panorama seams (B), abnormal sky patterns caused by extreme panoramic viewpoints (C), and local structural inconsistencies along roads and boundary objects (D).* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

generated images preserve task relevant visual information in practical applications.

Third, although our experiments show that the generated images maintain neighborhood-level structural separability comparable to real images (i.e., the between/within ratio remains largely consistent), we observe a phenomenon of proportional compression in the overall visual embedding distances, where both within-cell and between-cell distances decrease simultaneously. This suggests that while the model preserves the relative spatial relationships between neighborhoods, it compresses the absolute magnitude of visual differences. Although such proportional compression does not affect the relative separability structure, it may weaken the representation of fine-grained morphological variations, particularly at higher spatial resolutions or in detailed street view textures. Future work could address this limitation by incorporating structure-aware constraints during training, such as embedding-scale regularization or heterogeneity-preserving loss functions, which explicitly encourage models to maintain both spatial structural relationships and the absolute amplitude of visual variation. Such improvements may further enhance the model's ability to preserve micro-scale urban morphological characteristics in generated street view imagery.

In addition, the architectural choice of the generative module could be further explored in future work. In this study, the first stage of the framework adopts a conditional GAN to provide stable geometric alignment for satellite to street viewpoint transformation. Our experiments indicate that this design offers reliable structural consistency for large scale cross view mapping. Although more recent generative architectures such as SPADE based models, SEAN, and diffusion based variants have demonstrated strong capabilities in semantic image synthesis, our results suggest that a lightweight conditional GAN remains effective for geometry oriented viewpoint transformation, achieving comparable or better performance across several metrics, including PSNR (14.56), SSIM (0.4269), MS-SSIM (0.4382), and LPIPS (0.3453), which indicates more stable structural consistency and perceptual similarity (Table 10). Overall, these results indicate that SPADE-style generative architectures do not provide clear advantages over the conditional GAN design adopted in our framework for satellite-to-street viewpoint transformation.

Future work could therefore explore hybrid or alternative architectures, such as combining geometry aligned generative modules with diffusion-based refinement or semantic conditioned synthesis, to further enhance fine grained realism and object level detail while maintaining stable geometric alignment during viewpoint transformation.

Despite these limitations, our framework provides a strong foundation for equitable street view generation at scale. Future work will focus on improving fine-grained realism, reducing computational costs, and integrating with multimodal urban analytics to advance toward globally inclusive and sustainable urban intelligence.

8. Conclusion

In this work, we addressed the global imbalance of street view imagery coverage, a challenge that constrains urban research and risks reinforcing knowledge inequality between developed and developing regions. To mitigate this, we proposed GeoIdentity-Sat2Street, a geographic identity preserving framework that generates realistic street views from satellite images through polar transformation, ControlNet guidance, and location text-design. Our approach consistently outperforms strong baselines across pixel-level, structural, perceptual, and identity preserving metrics, while GPT-based evaluation confirms its semantic fidelity and visual realism. Beyond benchmarks, experiments on the NUS Global Streetscapes dataset and a case study in Kathmandu demonstrate its practical value in supplementing missing imagery and improving fairness in downstream applications such as road safety analysis. Looking forward, we plan to extend this framework to incorporate richer multimodal signals (e.g., auditory and textual context), enhance scalability for global deployment, and improve coverage in underrepresented regions such as Africa and other developing areas. We also aim to explore integration with urban planning and policy-making workflows. By bridging the gap in street view data availability, our work not only advances technical progress in cross-view generation but also contributes to more equitable, inclusive, and data-driven urban analytics at a global scale.

Data availability

The analytical framework, data, and codes for data analysis in this study are publicly available via the project repository: https://github.com/ai4city-hkust/GeoIdentity-Sat2Street.

Author agreement

I, Zongrong Li, certify that all authors have seen and approved the final version of the manuscript titled "Bridging Street View Coverage Disparities through Geographic Identity Preserving Generation from Satellite View." We warrant that this manuscript is the author's original work, has not been previously published, and is not under consideration for publication elsewhere.

CRediT authorship contribution statement

Zongrong Li: Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Fan Zhang:** Writing – review & editing, Supervision, Conceptualization. **Shaoqing Dai:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization. **Wufan Zhao:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 42401567), the Tertiary Education Scientific Research Project of Guangzhou Municipal Education Bureau (Grant No. 2024312159), the Guangzhou Municipal Science and Technology Bureau Program (Grant No. 2025A03J3640), and the Guangdong Provincial Project (Grant No. 2024QN11G095). It was also supported in part by National Natural Science Foundation of China (Grant No. 42371468), the Open Fund of the Technology Innovation Center for 3D Real Scene Construction and Urban Refined Governance, Ministry of Natural Resources of China (Grant No. 2024PF-1), a open grant from LREIS, Chinese Academy of Sciences, and the AI Research and Learning Base of Urban Culture, Guangdong Provincial Department of Education (Grant No. 2023WZJD008).

References

Alpherts, T., Ghebreab, S., Van Noord, N., 2025. Artifacts of Idiosyncracy in Global Street View Data. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, pp. 416–437. https://doi.org/10.1145/3715275.3732029.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: a review. Landsc. Urban Plan. 215, 104217. https://doi.org/10.1016/j.landurbplan.2021.104217.

Brooks, T., Holynski, A., Efros, A.A., 2023. Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 18392–18402. https://arxiv.org/abs/2211.09800.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818. https://arxiv.org/abs/1802.02611.

Chen, L., Lu, Y., Ye, Y., Xiao, Y., Yang, L., 2022. Examining the association between the built environment and pedestrian volume using street view images. Cities 127, 103734. https://doi.org/10.1016/j.cities.2022.103734.

Dai, S., Li, Y., Stein, A., Yang, S., Jia, P., 2024. Street view imagery-based built environment auditing tools: a systematic review. Int. J. Geogr. Inf. Sci. 38 (6), 1136–1157. https://doi.org/10.1080/13658816.2024.2336034.

Deng, B., Tucker, R., Li, Z., Guibas, L., Snavely, N., Wetzstein, G., 2024. Streetscapes: large-scale consistent street view generation using autoregressive video diffusion. In: ACM SIGGRAPH 2024 Conference Papers, pp. 1–11. https://doi.org/10.1145/3641519.3657513.

Fan, Z., Feng, C.-C., Biljecki, F., 2025. Coverage and bias of street view imagery in mapping the urban environment. Comput. Environ. Urban Syst. 117, 102253. https://doi.org/10.1016/j.compenvurbsys.2025.102253.

Fry, D., Mooney, S.J., Rodríguez, D.A., Caiaffa, W.T., Lovasi, G.S., 2020. Assessing google street view image availability in Latin American cities. J. Urban Health 97 (4), 552–560. https://doi.org/10.1007/s11524-019-00408-7.

Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2414–2423. https://doi.org/10.1109/CVPR.2016.265.

Guo, S., Jang, K.M., Duarte, F., Kang, Y., Ratti, C., 2025. Urban visual uniqueness: a landmark-free framework to quantify city's identity and distinctiveness from everyday scenes. Comput. Environ. Urban Syst. 122, 102351. https://doi.org/10.1016/j.compenvurbsys.2025.102351.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/cvpr.2016.90.

Hou, Y., Quintana, M., Khomiakov, M., Yap, W., Ouyang, J., Ito, K., Wang, Z., Zhao, T., Biljecki, F., 2024. Global Streetscapes — a comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. ISPRS J. Photogramm. Remote Sens. 215, 216–238. https://doi.org/10.1016/j.isprsjprs.2024.06.023.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.632.

Li, Z., Li, Z., Cui, Z., Qin, R., Pollefeys, M., Oswald, M.R., 2021. Sat2vid: Street-view panoramic video synthesis from a single satellite image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12436–12445. https://doi.org/10.1109/iccv48922.2021.01221.

Li, J., Li, D., Savarese, S., Hoi, S., 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. PMLR, pp. 19730–19742. https://arxiv.org/abs/2301.12597.

Li, W., He, J., Ye, J., Zhong, H., Zheng, Z., Huang, Z., Lin, D. and He, C., 2024. Crossviewdiff: A cross-view diffusion model for satellite-to-street view synthesis. arXiv preprint arXiv:2408.14765. https://arxiv.org/abs/2408.14765.

Li, Z., Su, Y., Wang, H., Zhao, W., 2025. BuildingView: Constructing urban building exteriors databases with Street View imagery and multimodal large language model. In: International Conference on Spatial Data and Intelligence. Springer Nature Singapore, Singapore, pp. 1–19.

Li, Z., Su, Y., Biljecki, F., Zhao, W., 2026b. BuildingMultiView:powering multi-scale building characterization with large language models and Multi-perspective imagery. Int. J. Appl. Earth Observ. Geoinf. 146, 105034. https://doi.org/10.1016/j.jag.2025.105034.

Liu, L., Li, H., 2019. Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5624–5633. https://doi.org/10.1109/cvpr.2019.00577.

Liu, Z., Zhang, E., Pan, S., Li, S., Long, Y., Witlox, F., 2025. Assessing urban emergency medical services accessibility for older adults considering ambulance trafficability using a deep learning approach. Sustain. Cities Soc. 132, 106804. https://doi.org/10.1016/j.scs.2025.106804.

Lu, X., Li, Z., Cui, Z., Oswald, M.R., Pollefeys, M., Qin, R., 2020. Geometry-aware satellite-to-ground image synthesis for urban areas. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 856–864. https://doi.org/10.1109/cvpr42600.2020.00094.

Miyato, T., Kataoka, T., Koyama, M. and Yoshida, Y., 2018. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957. https://doi.org/10.48550/arxiv.1802.05957.

Qian, M., Xiong, J., Xia, G.-S., Xue, N., 2023. Sat2Density: faithful density learning from satellite-ground image pairs. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3660–3669. https://doi.org/10.1109/iccv51070.2023.00341.

Regmi, K., Borji, A., 2018. Cross-view image synthesis using conditional gans. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 3501–3510. https://doi.org/10.1109/cvpr.2018.00369.

Ren, B., Tang, H. and Sebe, N., 2021. Cascaded cross mlp-mixer gans for cross-view image translation. arXiv preprint arXiv:2110.10183. https://arxiv.org/abs/2110.10183.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci. 9351, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Shi, Y., Yu, X., Campbell, D., Li, H., 2020. Where am I Looking At? Joint location and orientation estimation by cross-view matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4064–4072. https://doi.org/10.1109/cvpr42600.2020.00412.

Shi, Y., Campbell, D., Yu, X., Li, H., 2022. Geometry-Guided Street-View Panorama Synthesis from Satellite Imagery. IEEE Trans. Pattern Anal. Mach. Intell. 44 (12), 10009–10022. https://doi.org/10.1109/tpami.2022.3140750.

Sun, B., Liu, G., Yuan, Y., 2023. F3-Net: multiview scene matching for drone-based geo-localization. IEEE Trans. Geosci. Remote Sens. 61, 1–11. https://doi.org/10.1109/tgrs.2023.3278257.

Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L., 2021. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6488–6497. https://arxiv.org/abs/2103.06818.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2017. Instance Normalization: The Missing Ingredient for Fast Stylization. ArXiv.org. https://doi.org/10.48550/arXiv.1607.08022.

Umar, F., Amoah, J., Asamoah, M., Dzodzomenyo, M., Igwenagu, C., Okotto, L.-G., Okotto-Okotto, J., Shaw, P., Wright, J., 2023. On the potential of Google Street View for environmental waste quantification in urban Africa: an assessment of bias in spatial coverage. Sustainable Environ. 9 (1). https://doi.org/10.1080/27658511.2023.2251799.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612. https://doi.org/10.1109/tip.2003.819861.

Wang, J., Tan, H., Liao, B., Jiang, A., Fei, T., Huang, Q., Zhou, B., Tu, Z., Ye, S. and Kang, Y., 2025a. SounDiT: Geo-contextual soundscape-to-landscape generation. arXiv preprint arXiv:2505.12734. https://arxiv.org/abs/2505.12734.

Wang, L., Zhang, T., He, J., Kada, M., 2025b. Cross-platform complementarity: assessing the data quality and availability of google street view and Baidu street view. Trans. Urban Data, Sci., Technol. 4 (1), 22–47. https://doi.org/10.1177/27541231241311474.

Wei, J., Kan, Z., Liu, Z., Wang, Z., Chen, Y., 2025. Exploring spatio-temporal variations in nonlinear correlations between the built environment and metro-bike travel: evidence from Beijing. Transportmetrica B: Transport Dynamics 13 (1). https://doi.org/10.1080/21680566.2025.2541282.

Workman, S., Souvenir, R., Jacobs, N., 2015. Wide-area image geolocalization with aerial reference imagery. International Conference on Computer Vision. https://doi.org/10.1109/iccv.2015.451.

Wu, M., Waheed, A., Zhang, C., Abdul-Mageed, M., Aji, A.F., 2023. LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions. ArXiv.org. https://doi.org/10.48550/arXiv.2304.14402.

Xu, N., Qin, R., 2025. Satellite to GroundScape – large-scale consistent ground view generation from satellite views. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6068–6077. https://doi.org/10.1109/cvpr52734.2025.00569.

Yin, C., Peng, N., Li, Y., Shi, Y., Yang, S., Jia, P., 2023. A review on street view observations in support of the sustainable development goals. Int. J. Appl. Earth Obs. Geoinf. 117, 103205. https://doi.org/10.1016/j.jag.2023.103205.

Zhang, L., Agrawala, M., 2023. Adding Conditional Control to Text-to-Image Diffusion Models. ArXiv.org. https://arxiv.org/abs/2302.05543.

Zhang, Y., Li, S., Dong, R., Deng, H., Fu, X., Wang, C., Yu, T., Jia, T., Zhao, J., 2021. Quantifying physical and psychological perceptions of urban scenes using deep learning. Land Use Policy 111, 105762. https://doi.org/10.1016/j.landusepol.2021.105762.

Zhang, F., Salazar-Miranda, A., Duarte, F., Vale, L., Hack, G., Chen, M., Liu, Y., Batty, M., Ratti, C., 2024a. Urban visual intelligence: studying cities with artificial intelligence and street-level imagery. Ann. Am. Assoc. Geogr. 114 (5), 876–897. https://doi.org/10.1080/24694452.2024.2313515.

Zhang, J., Li, Y., Fukuda, T., Wang, B., 2024b. Urban Safety Perception Assessments via Integrating Multimodal Large Language Models with Street View Images. ArXiv.org. https://arxiv.org/abs/2407.19719.

Zhao, S., Chen, D., Chen, Y.-C., Bao, J., Hao, S., Yuan, L., Wong, K.-Y.K., 2023. Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models. ArXiv.org. https://doi.org/10.48550/arXiv.2305.16322k.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/iccv.2017.244.

Zhu, H., Nan, X., Yang, F., Bao, Z., 2023. Utilizing the green view index to improve the urban street greenery index system: a statistical study using road patterns and vegetation structures as entry points. Landsc. Urban Plan. 237, 104780. https://doi.org/10.1016/j.landurbplan.2023.104780.