OmniTex-Gen: Texturing LOD-1 Building Models via Panorama Synthesis from Virtual Street-View Depth

Xuming Zhang, Han Hu*, Liupeng Su, Junfan Wang, Yuhao Ye, and Qing Zhu

Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 611756, China

ABSTRACT

Satellite imagery has become a major data source for large-scale reconstruction of Level of Detail 1 (LOD-1) building models. However, satellite imagery provides limited views of façades, so these models typically use textures from generic libraries, limiting their visual realism. We present a framework that projects LOD-1 building models into virtual street-view depth maps to guide panorama synthesis, in order to generate realistic façade textures consistent with the building geometry. Specifically, a latent diffusion model uses these maps through ControlNet to constrain scene geometry. In addition, city-based textual conditioning provides shared streetscape cues, while satellite orthophoto features supply local appearance information through IP-Adapter. For training and evaluation, we construct OmniCityTex, containing 29,881 samples from 18 cities that pair LOD-1 building models, satellite orthophotos, virtual depth maps, and cleaned street-view panoramas. Experimental evaluations reveale that our method reduces the DINO-based structural distance and FID by 6.2% and 39.5%, respectively, and increases building-class intersection over union by 27.1% relative to the strongest baseline for each metric. Our model achieves lower perceptual distances and higher semantic agreement than the baselines when evaluating from four unseen cities. The dataset and model weights will be publicly available at https://vrlab.org.cn/~hanhu/projects/omnitex-gen/.

Keywords: LOD-1 Building Models; Satellite-Ground Synthesis; Façade Texture Mapping; Conditional Diffusion Model

1. Introduction

Reconstructing Level of Detail 1 (LOD-1) building models (OGC 2021) from 0.3–0.5 m satellite stereo imagery supports large-area urban modeling, as shown by the LOD-1 and LOD-2 building products in Maxar's Precision3D Telco Suite (Maxar 2024) and the satellite-based 3D reconstruction tasks within China's national 3D mapping program (3DRGLM) (Chen et al. 2025). However, the spatial resolution and viewing geometry of satellite imagery limit the recovery of façade details, often leaving reconstructed models without realistic textures (Yu et al. 2026). Generic texture libraries provide a practical way to add surface appearance to these simplified models (Zhang et al. 2022), supporting their use in urban digital twins and large-scale 3D visualization (Büyükdemircioğlu and Oude Elberink 2025). Recent advances in conditional diffusion models have improved control over visual style and the preservation of structural features during image generation (Cohen et al. 2025; Cheng et al. 2026). We generate visually plausible façade textures for LOD-1 building models while maintaining geometric consistency and regional appearance.

Satellite imagery mainly shows buildings from above (a nadir view), while façade appearance is observed at street level. Generating textures for satellite-derived building models therefore involves a cross-view problem (Xu et al. 2026), such as relating building footprints in overhead views to façades in street-view images. Prior studies show that the large viewpoint difference leaves little shared visual content for matching, as satellite images often reveal rooftops but few façade details (Toker et al. 2021; Lu et al. 2020). Without depth information, it is also difficult to determine which building outline corresponds to a visible façade, while projections based on a flat-ground assumption can distort buildings and produce incorrect matches (Song et al. 2023).

A second challenge is to generate plausible façade textures from LOD-1 models, which provide simplified building geometry without appearance information. Existing approaches synthesize street-view images using generative adversarial networks (GANs) trained on paired overhead and street-view images (Toker et al. 2021; Wu et al. 2023) or diffusion models that incorporate structural cues (Lee and Benes 2025). However, these methods mainly aim to generate complete street scenes. Façade textures for individual LOD-1 buildings should also be consistent with building height and number of floors, with elements such as windows appearing at appropriate sizes and spacing.

We therefore formulate the task as geometry- and style-conditioned street-view panorama synthesis from LOD-1 building geometry. We use LOD-1 building footprints and heights to represent building shapes from the target street-level viewpoint and constrain the geometry of the synthesized panorama. This target-domain representation specifies the spatial arrangement of buildings, while a conditional diffusion model generates visually plausible façade

*Corresponding author: Han Hu, han.hu@swjtu.edu.cn

appearance using city-based text prompts and satellite imagery.

Our framework first converts LOD-1 building footprints and height information into camera-aligned virtual street-view depth maps by adapting a multiplane projection strategy (Shi et al. 2022). These maps show the buildings' overall shapes without projecting satellite textures into the panoramas. A latent diffusion model (Rombach et al. 2022) uses these depth maps through ControlNet (Zhang et al. 2023) to constrain building positions and shapes. City names in a fixed text prompt provide shared streetscape cues, while satellite orthophoto features supply local appearance information through IP-Adapter (Ye et al. 2023). Training proceeds in two stages: we first fine-tune ControlNet, then freeze it and fine-tune selected IP-Adapter components. Both stages use an L_1 noise-prediction loss with the diffusion backbone frozen. With these geometric and style conditions, the model generates street-view panoramas that reflect both the LOD-1 building geometry and the regional appearance. We then project the generated panoramas into perspective views, manually extract façade textures, and automatically map them onto the corresponding LOD-1 building models.

To train this network, we construct a new dataset, OmniCityTex, using building footprints and height information from OpenStreetMap (OSM). The dataset pairs OSM-derived LOD-1 building models with satellite orthophotos, cleaned street-view panoramas, and corresponding virtual street-view depth maps. The training set contains 23,101 paired samples from 14 cities. The training panoramas require cleaning because vehicles, trees, and other foreground objects obscure façades and are absent from the depth maps. We use a large vision-language model (OpenAI 2026) to remove these occluding objects while preserving the camera viewpoint, building outlines, and visible façade details.

The main contributions are as follows.

- We construct OmniCityTex, a dataset of 29,881 samples from 18 cities across 15 countries worldwide. Each sample includes an LOD-1 building model, a satellite orthophoto, a cleaned street-view panorama, and a corresponding virtual street-view depth map.
- We develop a method that projects LOD-1 building footprints and heights into virtual street-view depth maps and uses these maps to constrain panorama generation. The depth maps constrain building positions and shapes, while city-based textual conditioning and satellite orthophoto visual conditioning guide façade appearance.

2. Related Work

2.1. Geometry-Conditioned Street-View Synthesis

Image-based 3D reconstruction uses multiview stereo and neural scene representations (Zhu et al. 2026), with multi-camera tiling enabling neural radiance fields to process large aerial images (Xu et al. 2024). After cross-view geometry has been aligned, a synthesis model must infer ground-level appearance that top-down data capture poorly or miss entirely. Early generative methods formulate this task as conditional image translation. Conditional generative adversarial networks (GANs) generate target-view RGB images together with semantic maps (Regmi and Borji 2018). Panorama-oriented GANs use adversarial feedback and pixel–semantic alignment to refine both predictions (Wu et al. 2023). Semantic maps describe object classes and spatial layout, but they do not specify façade materials or architectural details. Latent diffusion models provide a flexible generative prior through iterative denoising in a compressed image space (Rombach et al. 2022). ControlNet introduces spatial conditions, such as depth and edges, into a pretrained diffusion model (Zhang et al. 2023), while low-rank adaptation (LoRA) supports parameter-efficient model adaptation (Hu et al. 2021).

Recent diffusion methods combine structural, pose, and appearance conditions for cross-view synthesis. Structure-focused methods construct separate structural and texture controls and inject them through enhanced cross-view attention (Li et al. 2024), establish explicit correspondences between satellite and street-view features (Lin et al. 2025), or combine aerial image features with estimated height and semantic embeddings (Lee and Benes 2025). Other methods add explicit controls over pose and regional appearance. Controllable Satellite-to-Street-View Synthesis (CS2S) iteratively adjusts pose during denoising and uses text prompts to control environmental conditions (Ze et al. 2025). Beyond its geometric component, Geospecific View Generation combines projected satellite cues with a learnable geographic prior and LoRA to model regional appearance (Xu and Qin 2024). For view sequences, Satellite to GroundScape combines denoising conditioned on satellite imagery with satellite-temporal denoising (Xu and Qin 2025), whereas SatDreamer360 uses a tri-plane representation, ray-based pixel attention, and panoramic epipolar-constrained attention (Ze et al. 2026). These methods mainly synthesize complete street scenes or view sequences rather than rectified façade textures ready for mapping onto LOD-1 building surfaces.

2.2. Façade Texture Generation and LOD-1 Texture Mapping

Façade texture mapping requires reliable correspondence between source imagery and 3D building surfaces.

Classical pipelines therefore combine source view selection with the handling of occlusion and radiometric differences. For individual façades, multiview coplanar extraction establishes homography-constrained correspondences from aerial and terrestrial imagery and then refines façade boundaries with geometric topology (He et al. 2022). At scene scale, loopy belief propagation supports view selection, multiview fusion, and global color consistency (Ling and Qin 2023). Semantic labels and object-type priors further guide view selection, reduce moving-object artifacts, and improve texture-boundary consistency (Liu et al. 2025). Street-view panoramas capture vertical façades directly from ground-level viewpoints. Adaptive viewpoint selection from ground panoramic videos improves geometric and texture reconstruction (Zhang et al. 2024). Cross-view surveying combines aerial and ground observations to improve building completeness (Tang et al. 2026). Panorama-based methods use semantic object matching and 3D inpainting to map sparse observations onto simplified urban geometry (Park et al. 2021), or they rectify panoramas against low-detail building models before façade segmentation and LOD3 reconstruction (Tang et al. 2025). At city scale, ray-traced visibility analysis and texture-atlas construction improve image selection and texture organization (Büyükdemircioğlu and Oude Elberink 2025). These pipelines still require imagery of the target scene.

Generative methods complete or restore façade appearance when observed textures are incomplete or damaged. TwinTex aligns and stitches selected multiview photographs on piecewise-planar architectural geometry before using diffusion to inpaint missing texture regions (Xiong et al. 2023). Semantic image translation instead synthesizes textures from structured façade labels and reference style information while maintaining architectural regularity (Shang et al. 2024). Diffusion-based restoration uses multimodal prompts to address several types of texture defects (Liu et al. 2026). FACTS requires no additional training and combines semantic inpainting with text prompts, multiscale edge constraints, and illumination harmonization to restore existing LOD-1 and LOD-2 textures (Cheng et al. 2026). Most reviewed studies address only one stage of the workflow. Few integrate façade synthesis from virtual street-view depth maps with LOD-1 texture mapping when imagery of the target building is unavailable.

3. Construction of the OmniCityTex Dataset

We construct OmniCityTex by geographically pairing OpenStreetMap (OSM) building data (Haklay and Weber 2008) with Google Street View panoramas to train and evaluate street-view panorama generation for LOD-1 building texturing. OSM provides the building footprints and height-related attributes used to construct LOD-1 geometry, while Google Street View captures the corresponding street-level appearance. Each sample pairs an LOD-1 building model and a satellite orthophoto with a cleaned street-view panorama and a virtual street-view depth map. We retrieve panoramas and their geographic coordinates through the Street View endpoints of the Google Map Tiles API (Google 2026), then use each panorama center to obtain the corresponding OSM building data and satellite orthophoto. The OSM building footprints and heights define the LOD-1 models, which we project into virtual depth maps aligned with the panoramas. We also remove foreground objects from the panoramas to prepare the training targets. The following subsections describe the OSM data preprocessing, geometric projection, and panorama cleaning used to build the dataset.

3.1. Virtual Camera Configuration

Camera position and height. We place the virtual camera at the panorama's geographic location, represented by the center (u_0, v_0) of a local top-down OSM grid. We place the camera at a fixed height of $h_c = 2.0\,\mathrm{m}$ above the local ground plane to approximate the height of the vehicle-mounted Street View camera.

Road-based initialization. We set the virtual camera pitch and roll to zero for the equirectangular panoramas, placing the horizon at the panorama equator. We use the local tangent of the nearest OSM road segment to initialize the horizontal orientation. Let R denote one direction along this tangent and β its compass bearing, measured clockwise from north N. We initially place R at the panorama center column, corresponding to 180° on the horizontal image coordinate (Fig. 1). The angular offset γ between R and the panorama's central viewing direction is therefore initialized to 0° (Fig. 2).

3.2. Virtual Street-View Depth Map Generation

Using the camera configuration in Section 3.1, we render a virtual depth map from a height-indexed occupancy volume of OSM buildings.

Height map construction. We extract building footprints from OSM and assign each building a height using its `height` attribute. If this attribute is missing, we estimate the height from `building:levels` using a fixed floor height, or use a category-based default when neither attribute is available. We rasterize the footprints on the local OSM grid at a resolution of 1280×1280, assigning building heights to cells inside footprints and zero elsewhere. The grid covers the same area as the corresponding zoom-level-20 satellite orthophoto; its size in metres varies with the local latitude.

Multiplane occupancy representation. Inspired by multiplane scene representations (Shi et al. 2022), we sample the height map at a sequence of horizontal elevations.

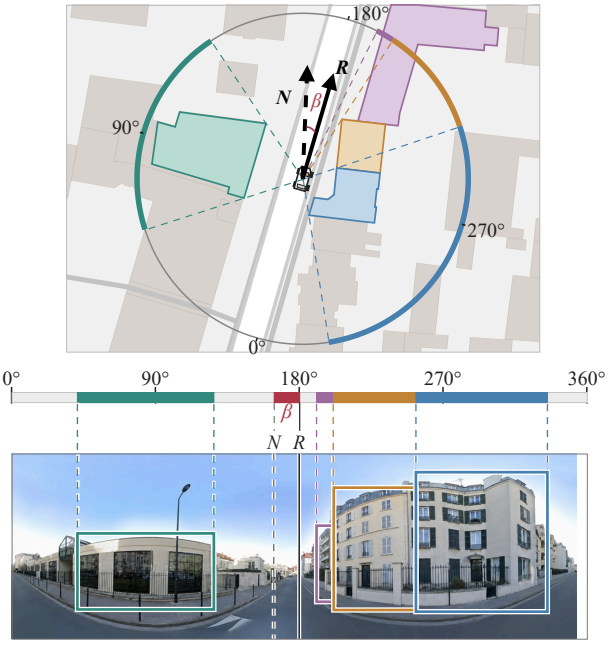


FIG. **1.** Angular correspondence between OSM and street-view imagery. Matching colors link building footprints and their angular extents to the corresponding regions in the panorama.

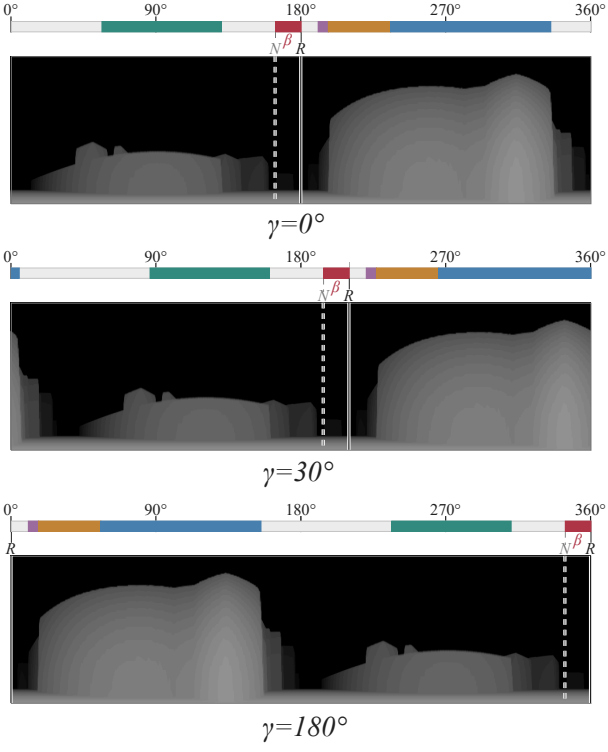


FIG. **2.** Depth-map alignment by horizontal circular shifts. Adjusting γ, the angular offset between R and the panorama's central viewing direction, circularly shifts the depth map.

With the camera center at $z = 0$, we use 128 uniformly spaced planes from the ground at $z = -h_c$ to $z = 80\,\mathrm{m}$, with a vertical interval of approximately 0.65 m. Each plane retains the horizontal resolution of the height map. Each layer is a binary occupancy map: a cell is occupied if it lies within a building footprint and the building reaches that layer's elevation; otherwise, it is empty.

Virtual panoramic projection. We project occupied samples into panoramic coordinates (Fig. 3). Let (u, v) denote the horizontal grid position of an occupied sample, z its camera-relative elevation, and s the number of grid units per metre. With the grid axes aligned to the panorama orientation, the sample projects to the azimuth ϕ and polar angle θ as

$$\theta = \begin{cases} \operatorname{atan2}\left(\sqrt{(v-v_0)^2 + (u-u_0)^2},\, sz\right), & z \neq 0, \\ \pi/2, & z = 0, \end{cases}$$
$$\phi = \begin{cases} \operatorname{atan2}(v - v_0, u - u_0), & u \neq u_0, \\ (\pi/2)\,\operatorname{sign}(v - v_0), & u = u_0. \end{cases} \tag{1}$$

Here, θ is measured from the upward vertical axis, so $\theta = \pi/2$ corresponds to the horizon.

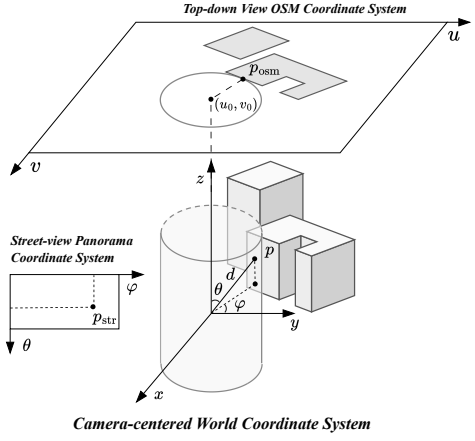
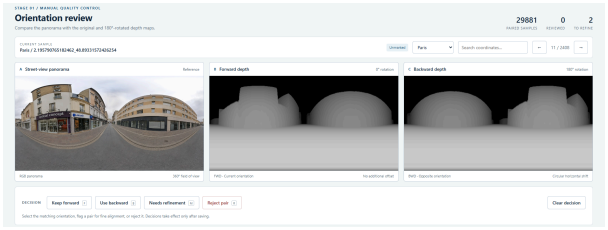
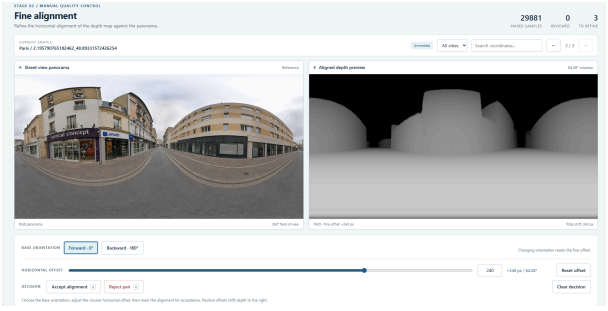


FIG. **3.** Geometric mapping from OSM to the panorama. A scene point projects to the panorama through its camera-centered azimuth ϕ and polar angle θ.

Visibility and depth assignment. To resolve occlusions, we sample the occupancy layers along camera-centered concentric circles and stack the unrolled samples into a cylindrical volume. Each panorama pixel defines a viewing ray through this volume according to Eq. (1). Far-to-near alpha compositing with occupancy-derived opacity allows nearby surfaces to occlude distant ones; for binary, opaque occupancy, the nearest occupied sample is visible. A horizontal ground plane terminates downward rays that do not intersect a building.



(a)



(b)

FIG. **4.** Depth-map alignment interfaces: (a) road-direction selection; (b) horizontal shift adjustment.

For the visible surface sample at (u, v, z), its depth is the Euclidean camera-to-sample distance.

$$d(u, v, z) = \sqrt{\frac{(v - v_0)^2 + (u - u_0)^2}{s^2} + z^2}. \qquad (2)$$

Here, division by s converts horizontal grid offsets into the same metric units as z. We assign this distance to the panorama pixel and apply logarithmic normalization, with higher intensities indicating nearer surfaces. This depth map serves as the geometric condition for panorama generation.

3.3. Panorama Curation and Dataset Assembly

Direction selection and residual alignment. To resolve the forward–backward ambiguity along the road, we generate two candidate depth maps with orientations differing by 180°. We developed two web interfaces to interactively select the direction and correct remaining alignment errors when preparing the dataset. In the first part (Fig. 4(a)), we compare both candidates with the street-view panorama and select the matching orientation. Most panoramas require only this direction selection. For the few cases with remaining alignment errors, we use the second interface (Fig. 4(b)) to adjust γ and circularly shift the depth map until corresponding building boundaries align, before accepting the pair.

Panorama cleaning. Real-world street-view panoramas often contain foreground objects, including vehicles, pedestrians, and vegetation, that obscure building façades. These objects are absent from the virtual street-view depth maps. Their presence in the target panoramas therefore creates inconsistencies during training and reduces the visible façade area available for texture extraction and mapping.

We use a large vision-language model (OpenAI 2026) to automatically remove foreground occlusions from the panoramas. The editing prompt instructs the model to remove foreground occluding objects while retaining the camera viewpoint, panoramic geometry, building outlines, and visible façade appearance. Fig. 5 shows clearer views of the building façades after foreground objects are removed. The prompt is shown in the box below.

> Edit this panoramic street-view image for a building-façade dataset.
>
> **Primary edits:**
> • Remove vehicles, pedestrians, trees, and temporary objects.
> • Preserve architecture, textures, lighting, and panorama geometry.
> • Do not change the viewpoint or layout.
>
> **Hard constraints:**
> • Keep the same street-view panorama/equirectangular geometry.
> • Keep the same 1024 × 512 panorama framing when possible.
> • Do not crop, zoom, rotate, straighten, or change the viewpoint.
> • Do not change the street layout, building silhouettes, architecture, façade textures, windows, road, sky, lighting, or overall photographic style.
> • Remove only occluding vehicles, pedestrians, trees, and temporary objects, and repair the revealed façade and street regions naturally.
> • Do not add text, logos, watermarks, new signs, new buildings, or stylized effects.

We pair each cleaned panorama with the LOD-1 building model, satellite orthophoto, and virtual street-view depth map from the same location. The building models and depth maps are constructed using the procedure in Section 3.2. The resulting data form OmniCityTex, a dataset for generating textures for LOD-1 building models. OmniCityTex contains 29,881 samples from 18 cities across 15 countries. Fig. 6 summarizes their geographic distribution. Table 1 summarizes the data by city, including panorama counts, unique building counts, spatial coverage, and mean building height.

OmniCityTex includes LOD-1 building models, satellite orthophotos, virtual street-view depth maps, and cleaned street-view panoramas. These four data components support façade texture generation and future research. As shown in Fig. 7, each example links the four components from the same location.

4. Panorama Synthesis from Virtual Street-View Depth

We synthesize street-view panoramas using a latent diffusion model conditioned on city-based text prompts, satellite orthophotos, and virtual street-view depth maps (Fig. 8). Text prompts provide shared scene semantics, satellite orthophotos supply local appearance cues through IP-Adapter, and depth maps constrain scene geometry through ControlNet. We extract façade textures from the generated panoramas and automatically map them onto LOD-1 building surfaces.



FIG. **5.** Foreground-occlusion cleaning in OmniCityTex. The cleaning step removes visible foreground objects and reveals more of the building façades.

TABLE **1.** City-level statistics of OmniCityTex.

City	Panoramas	Unique buildings	Coverage (km^2)	Mean height (m)
Hong Kong	2,683	8,316	6.04	35.25
Berlin	2,624	18,897	12.14	14.73
Paris	2,408	40,075	13.14	16.64
Dubai	2,101	8,046	11.73	17.19
Rome	2,076	4,124	3.91	15.59
London	2,055	21,826	8.02	13.57
New York	1,889	12,701	5.41	25.04
Vienna	1,807	5,627	4.48	15.60
Taipei	1,738	5,676	5.77	19.89
Tokyo	1,721	32,174	9.92	15.51
Brasilia	1,652	38,198	11.74	5.61
Los Angeles	1,615	12,272	10.02	9.08
Chicago	1,583	12,080	5.79	13.34
Ulaanbaatar	1,503	5,133	8.67	17.39
Johannesburg	762	3,816	3.90	18.92
Moscow	685	1,915	3.78	23.47
Canberra	603	1,612	4.17	14.21
Mexico City	376	3,950	2.71	18.17
18 cities	**29,881**	**236,438**	**131.35**	**14.98**

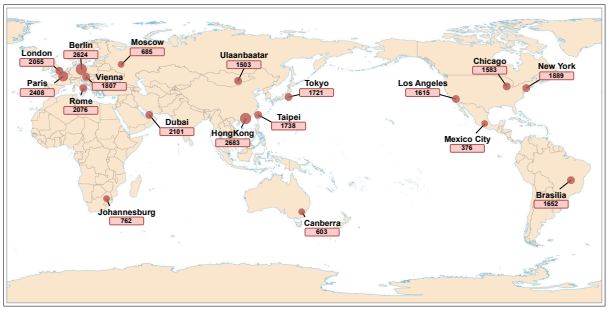


FIG. **6.** Geographic distribution of OmniCityTex. The dataset contains 29,881 samples from 18 cities across 15 countries.

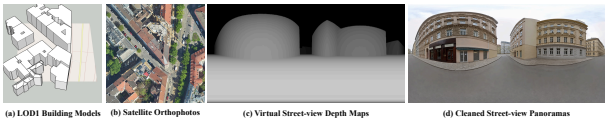


FIG. **7.** Multimodal contents of OmniCityTex. Each sample links an LOD-1 building model and satellite orthophoto with a virtual street-view depth map and a cleaned street-view panorama from the same location.

4.1. Textual Conditioning from City-Based Style

Street scenes vary across cities in façade design, storefronts, and building density. For example, dense streets with many ground-floor storefronts in Hong Kong and streets lined with classical façades in Paris have different visual styles. We use city names as textual cues for these recurring streetscape styles.

We use the fixed prompt structure "*A photorealistic 360-degree street-view panorama in [city]*," retaining the shared scene description and varying the city term to represent the target streetscape style. Fig. 9 shows that city-based textual conditioning produces different streetscape styles while retaining the geometric layout specified by the same virtual depth map.

4.2. Visual Conditioning from Satellite Orthophoto

City-based textual conditioning provides shared scene semantics but does not fully describe appearance differences between locations within the same city. With the same text prompt, different virtual depth maps may produce panoramas with different geometric layouts but similar visual styles. We use the corresponding satellite orthophoto to add local appearance cues to the textual condition.

We use a pretrained IP-Adapter (Ye et al. 2023) for satellite orthophoto conditioning. Its frozen CLIP image encoder (Radford et al. 2021) maps the full satellite orthophoto S_i, resized to the encoder's input dimensions without centre cropping, to a global image embedding. The adapter's image projection, consisting of a linear layer

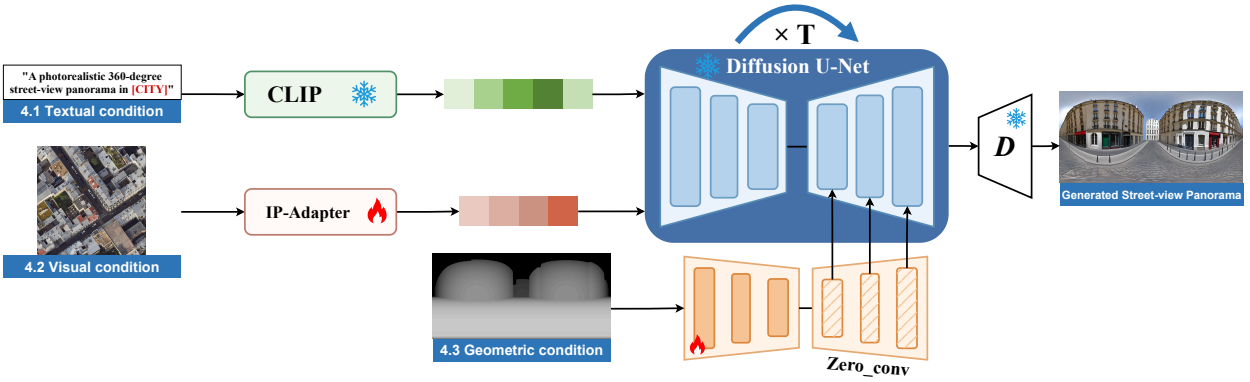


FIG. **8.** Panorama synthesis framework. City-based text prompts, satellite orthophoto features, and virtual depth maps guide panorama generation through text cross-attention, IP-Adapter, and ControlNet, respectively.

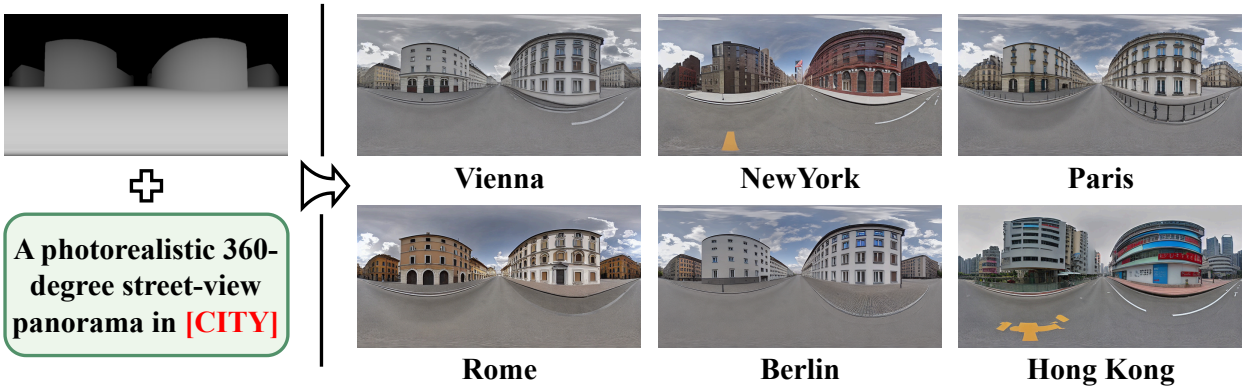


FIG. **9.** Effect of city-based textual conditioning. We vary the city name in the text prompt while keeping the virtual depth map fixed and disabling satellite orthophoto conditioning.

and layer normalization, converts this embedding into four visual tokens $\mathbf{c}_i^{\mathrm{v}}$.

We enable image cross-attention only in the second upsampling block of the U-Net decoder. At these layers, the text and image branches share queries derived from the U-Net features and attend separately to the text and visual tokens. Their outputs are summed to condition the denoising features. We fine-tune the image projection and the key and value projections of these image-attention layers, while the original U-Net parameters remain frozen.

4.3. Geometric Conditioning from Depth Map

We finetune a pretrained ControlNet (Zhang et al. 2023) to condition panorama synthesis on the virtual street-view depth map D_i constructed in Section 3.2. Its multiscale features pass through zero-convolution connections and are added to the U-Net's corresponding skip and middle-block features to constrain building positions and shapes.

Each training sample contains a cleaned panorama I_i, satellite orthophoto S_i, virtual depth map D_i, and city label. The frozen VAE encoder maps I_i to a latent representation $\mathbf{z}_{i,0}$. At a randomly sampled diffusion step t, we add Gaussian noise:

$$\mathbf{z}_{i,t} = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_{i,0} + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\alpha}_t$ is the cumulative product of the noise-schedule coefficients. The U-Net combines the noisy latent and timestep with the geometric residuals and the textual and visual conditions to predict the added noise:

$$\hat{\boldsymbol{\epsilon}}_i = \epsilon_\theta(\mathbf{z}_{i,t}, t, \mathbf{c}_i^{\mathrm{t}}, \mathbf{c}_i^{\mathrm{v}}, D_i). \tag{4}$$

We minimize the mean absolute error between the sampled and predicted noise over all latent channels and spatial positions:

$$\mathcal{L}_{L_1} = \frac{1}{CHW} \sum_{c,h,w} \left| \boldsymbol{\epsilon}_{c,h,w} - \hat{\boldsymbol{\epsilon}}_{i,c,h,w} \right|, \tag{5}$$

where C, H, and W are the channel count, height, and width of the latent representation.

We optimize this loss in two stages. The first stage uses depth and text conditions to fine-tune all ControlNet parameters. In the second stage, we freeze ControlNet and introduce satellite orthophoto conditioning, fine-tuning the IP-Adapter image projection and the image-attention key and value projections in the second U-Net upsampling block. The original U-Net parameters, VAE, and CLIP text and image encoders remain frozen.

4.4. Panorama Generation and Façade Texturing

At inference, we initialize a latent with Gaussian noise and iteratively denoise it using the satellite orthophoto, virtual depth map, and city-based text prompt. The VAE decoder converts the final latent into a street-view panorama.

We project the equirectangular panorama into perspective views using selected viewing directions and fields of view (Fig. 10). For each pixel in the perspective view, we determine its viewing direction and sample the corresponding location in the equirectangular panorama. We keep the virtual camera level and shift the viewing window vertically to frame the façade.

The projected views may contain multiple building façades, and distant façades or those extending high above the viewpoint may retain distortion. We therefore developed a web interface to extract individual façade textures through manual selection of four corner points (Fig. 11). A planar homography transforms the selected quadrilateral into a rectangular image, which is exported as a façade texture. The exported textures are automatically associated with the corresponding LOD-1 building surfaces and mapped onto them.

5. Experiments

5.1. Implementation Details

Dataset splits. We select the test and generalization test samples from OmniCityTex using geographic farthest-point sampling with a random seed of 42. Within each city, each new sample maximizes the distance to its nearest previously selected sample. We exclude the four cities with the fewest samples—Canberra, Johannesburg, Mexico City, and Moscow—from training and reserve them for generalization testing. From the remaining 14 cities, we select 20 samples per city, giving 280 test samples.

After excluding samples within 200 m of any test location, we use the remaining 23,101 paired samples from these 14 cities for training. This spatial separation reduces similarity between the training and test sets.

For generalization testing, we select 50 samples from each of the four unseen cities, giving 200 samples.

Training settings. We use Stable Diffusion v1.5 (Rombach et al. 2022) with pretrained depth-ControlNet and SD1.5 IP-Adapter weights. The training panoramas and depth maps have a resolution of 1024×512 pixels. We

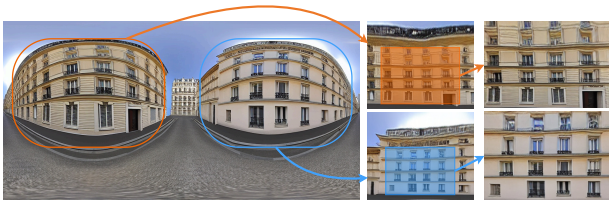


FIG. **10.** Panorama projection and texture extraction. The generated panorama is projected into perspective views, from which selected façade regions are transformed into rectangular textures.

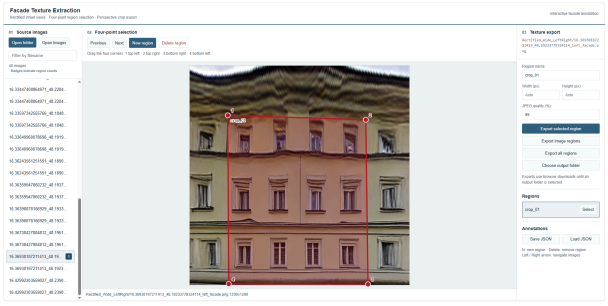


FIG. **11.** Four-point façade texture extraction interface. We mark four corners of a façade region in a perspective view and export the rectified texture.

specify the city in the U-Net prompt "A photorealistic 360-degree street-view panorama in [city]."

Training proceeds in two stages, first training depth-guided synthesis and then adding satellite orthophoto conditioning. In the first stage, we optimize ControlNet for 12,000 steps at a learning rate of 10^{-5}, using depth and text conditions. In the second stage, we freeze ControlNet and optimize the trainable components of IP-Adapter for 8,000 steps at a learning rate of 5×10^{-6}. These components are the image projection and the image-attention key and value projections in the second U-Net upsampling block. The pretrained U-Net parameters, VAE, and CLIP text and image encoders remain frozen throughout training.

Both stages use AdamW and an L_1 noise-prediction loss over the full latent map. Learning rates follow cosine schedules after 500 warm-up steps in the first stage and 400 in the second. We use bf16 mixed precision and gradient checkpointing on four GPUs, with one sample per GPU and four gradient-accumulation steps, giving an effective batch size of 16. To support classifier-free guidance (Ho and Salimans 2022), we replace the U-Net prompt with an empty string with probability 0.1.

Inference settings. We generate 1024×512 panoramas using Denoising Diffusion Implicit Models (DDIM) (Song et al. 2020) with 30 denoising steps and a classifier-free guidance scale of 7.5. The ControlNet and IP-Adapter conditioning scales are both 1.0. Inference uses fp16 precision, a fixed seed of 42, and an empty negative prompt.

TABLE **2.** Quantitative comparison on the test set. We evaluate all four methods on the same 280 samples.

Method	Pixel level		Structural level		Perceptual level			Semantic level		
	RMSE↓	PSNR↑	I_s ↑	D_s ↓	P_{squeeze} ↓	P_{alex} ↓	FID↓	I_B ↑	Acc.↑	mIoU↑
Sat2Density++	**56.113**	**13.300**	0.0530	0.1540	0.5333	0.6801	166.85	0.4503	0.6338	0.1460
Sat3DGen	56.746	13.240	0.0570	0.1433	0.5090	0.6682	132.17	0.4814	0.6813	0.1600
CS2S	62.082	12.431	0.0459	0.1507	0.5303	0.7105	130.12	0.3845	0.6475	0.1465
Ours	62.818	12.225	**0.0681**	**0.1344**	**0.4202**	**0.6059**	**78.73**	**0.6120**	**0.8258**	**0.1982**

We use the same trained model and city-prompt template for the 14 training cities and the four unseen cities.

5.2. Baselines and Metrics

Baselines. We compare our method with three approaches for synthesizing street-view imagery from satellite images. Sat2Density++ uses a satellite-conditioned neural radiance field to render view-consistent panoramas at specified camera locations (Qian et al. 2026a). Sat3DGen generates a structured, renderable street-level 3D scene from a single satellite image using a geometry-first strategy (Qian et al. 2026b). Controllable Satellite-to-Street-View Synthesis (CS2S) combines diffusion-based cross-view conditioning with iterative homography adjustment for pose alignment, and uses text prompts for zero-shot control of illumination and weather (Ze et al. 2025). We use the same OmniCityTex test samples for all methods, preprocess inputs according to each method's requirements, and evaluate outputs against the corresponding ground-truth panoramas.

Metrics. We use RMSE and PSNR to measure pixel-level similarity. These metrics measure pixel differences but do not fully capture the quality of cross-view synthesis (Regmi and Borji 2018; Lu et al. 2020). Following previous studies (Xu and Qin 2024; Ze et al. 2025), we also evaluate structural, perceptual, and semantic similarity.

For structural evaluation, we compute the mean intersection over union (IoU) of Canny edge maps (Canny 1986) from synthesized and ground-truth panoramas, denoted by I_s. We also measure the distance between their DINO-based structural self-similarity representations, denoted by D_s (Tumanyan et al. 2022). Perceptual similarity is evaluated using FID (Heusel et al. 2017) and LPIPS (Zhang et al. 2018); P_{alex} and P_{squeeze} denote LPIPS scores with pretrained AlexNet (Krizhevsky et al. 2012) and SqueezeNet (Iandola et al. 2016) backbones, respectively.

For semantic evaluation, we use OneFormer (Jain et al. 2023) pretrained on ADE20K (Zhou et al. 2017) to label synthesized and ground-truth panoramas. We report pixel accuracy (Acc.) and mean IoU (mIoU), together with building-class IoU I_B to assess agreement in the regions used for façade texturing.

5.3. Comparison with State-of-the-Art Methods

Table 2 reports the comparison with Sat2Density++, Sat3DGen, and CS2S on the 280 OmniCityTex test samples, and Fig. 12 shows eight examples.

Structural-level similarity. Our method has the highest edge IoU I_s and lowest DINO-based structural distance D_s, with a 19.6% increase and a 6.2% reduction, respectively, relative to Sat3DGen. In the second and third rows of Fig. 12, our results follow the main building outlines and street openings specified by the depth maps.

Perceptual-level similarity. Our method has the lowest FID and both LPIPS scores. Its FID of 78.73 is 39.5% lower than the lowest baseline value of 130.12, obtained by CS2S. The fifth row of Fig. 12 illustrates differences in façade detail: window and door boundaries are distinct in our output, while the baseline outputs contain blurred or distorted façades.

Semantic-level similarity. Our method has the highest pixel accuracy, mIoU, and building-class IoU I_B. Its I_B is 27.1% higher than that of Sat3DGen, indicating closer agreement with the reference building masks. In the third row of Fig. 12, our generated building façades are visible without the vehicles and roadside trees present in the baseline outputs.

Pixel-level similarity. Our method ranks fourth on both RMSE and PSNR. Fig. 12 shows differences from the ground truth in façade colour, window arrangement, and road markings. For example, the seventh row contains yellow façades and decoration around the windows in our output, while the reference façades are mostly white with simpler details.

5.4. Ablation of Geometric and Visual Conditioning

We disable satellite orthophoto conditioning, ControlNet, or both during inference, using the same trained model, city-based text prompts, and random seed. Table 3 reports results on the 280 test samples, and Fig. 13 shows qualitative examples.

Geometric conditioning. Our full model has lower RMSE, FID, and both LPIPS scores and higher PSNR than the variant without ControlNet (Table 3). It also increases I_s, pixel accuracy, and mIoU, reduces D_s from 0.2054 to 0.1344, and increases I_B from 0.2461 to 0.6120.

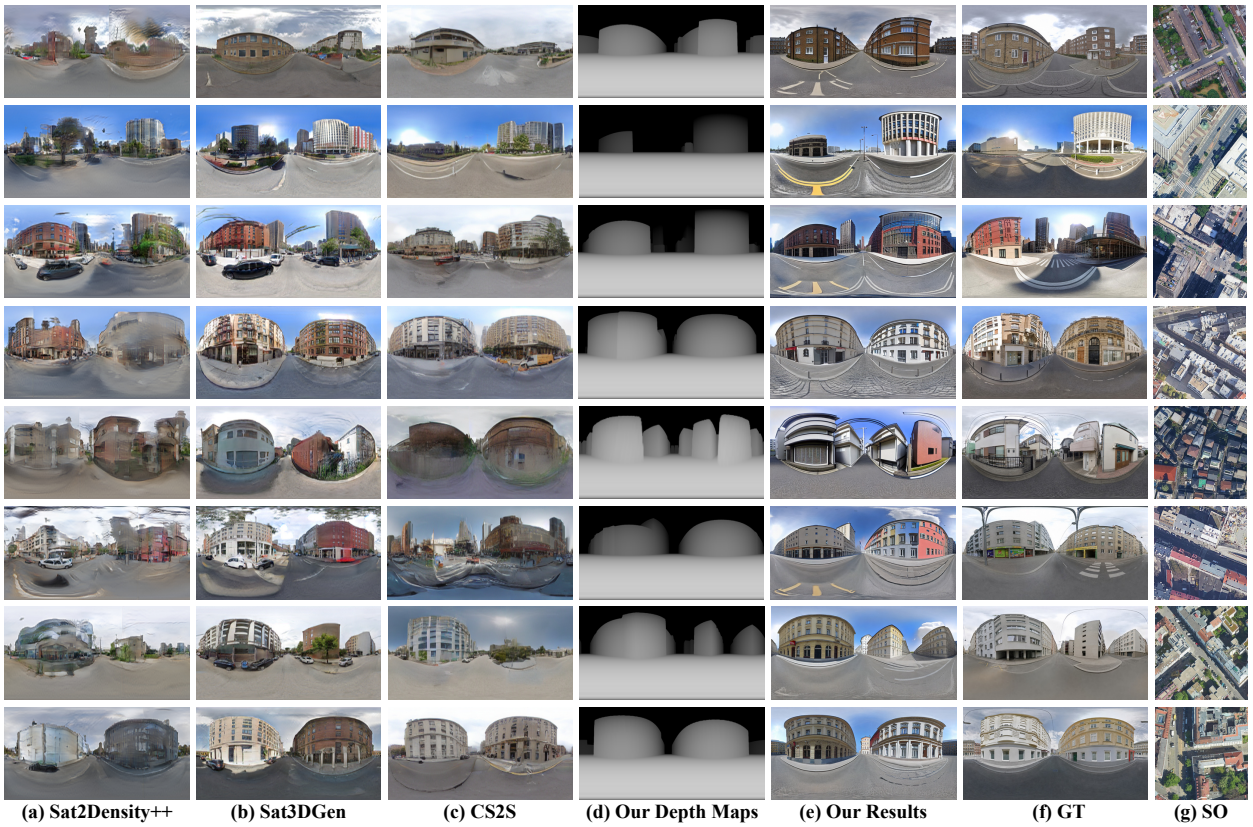


FIG. **12.** Qualitative comparison on the test set. Results from Sat2Density++, Sat3DGen, CS2S, and our method are shown alongside virtual depth maps, ground-truth panoramas, and satellite orthophoto (SO) inputs.

TABLE **3.** Quantitative conditioning ablation. We disable satellite orthophoto (SO) conditioning, ControlNet, or both during inference and evaluate each configuration on the same 280 test samples.

Method	**Pixel level**		**Structural level**		**Perceptual level**			**Semantic level**		
	RMSE↓	PSNR↑	I_s ↑	D_s ↓	P_{squeeze} ↓	P_{alex} ↓	FID↓	I_B ↑	Acc.↑	mIoU↑
w/o SO & ControlNet	70.108	11.276	0.0609	0.2083	0.5617	0.7332	242.10	0.2313	0.3496	0.0462
w/o ControlNet	71.503	11.122	0.0598	0.2054	0.5617	0.7421	139.45	0.2461	0.3804	0.0519
w/o SO	**61.987**	**12.343**	0.0672	0.1351	**0.4133**	**0.5882**	81.43	0.6066	**0.8281**	0.1945
Ours	62.818	12.225	**0.0681**	**0.1344**	0.4202	0.6059	**78.73**	**0.6120**	0.8258	**0.1982**

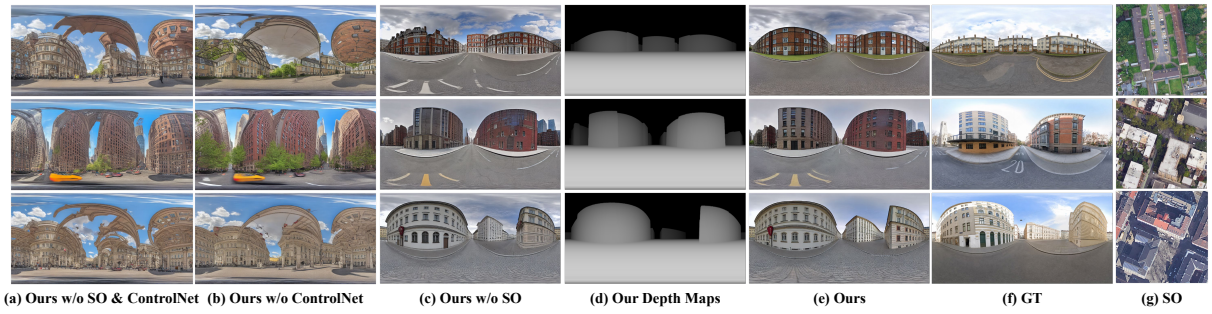


FIG. **13.** Qualitative conditioning ablation. We compare the full model with results obtained by disabling satellite orthophoto (SO) conditioning, ControlNet, or both during inference.

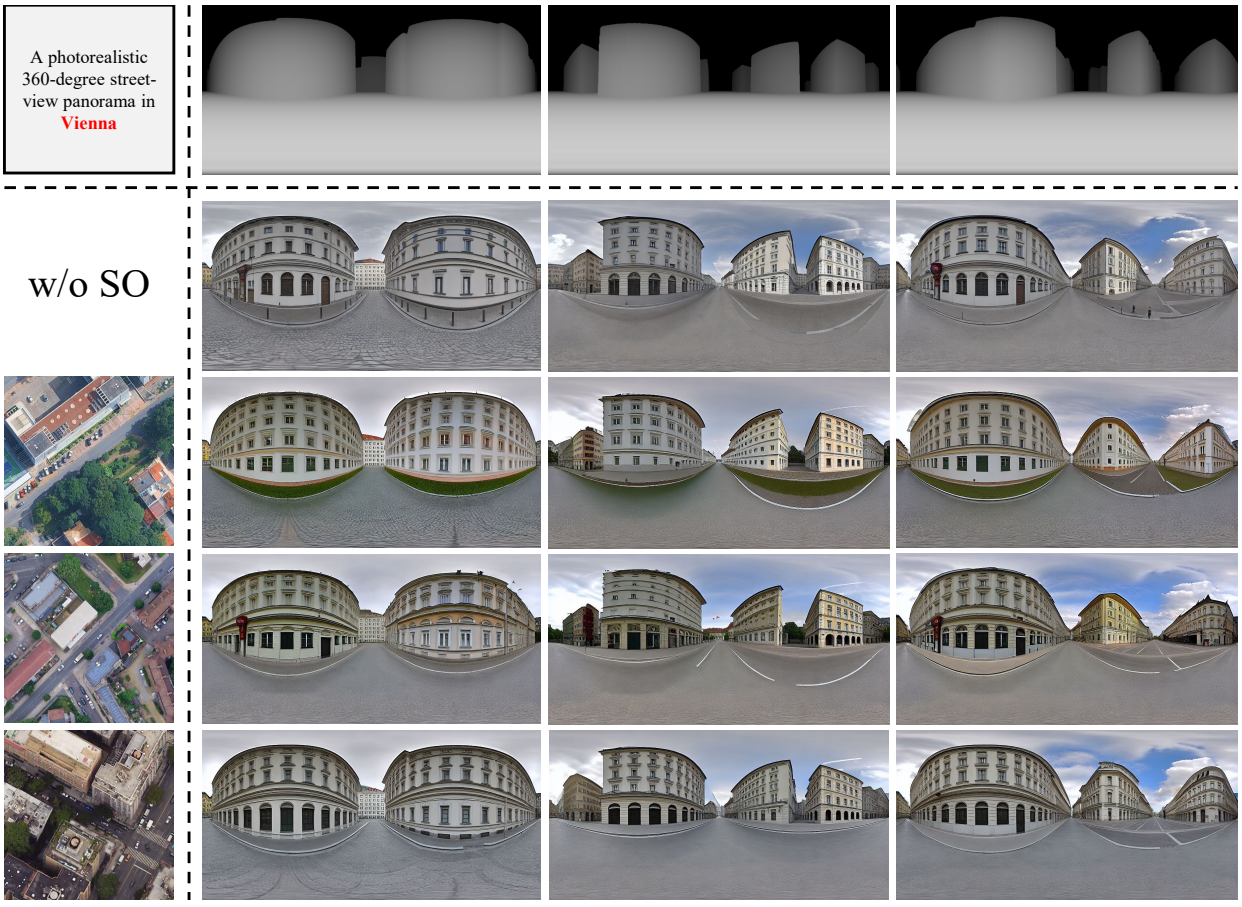


FIG. **14.** Effect of visual conditioning. The top row shows the depth maps used with the Vienna text prompt; the first panorama row omits satellite orthophoto (SO) conditioning, and the remaining rows use the SO shown at left.

Fig. 13(b) shows distorted building surfaces and street layouts that differ from the depth maps. Disabling both satellite orthophoto conditioning and ControlNet produces similar geometric distortions (Fig. 13(a)), with $D_s = 0.2083$, and raises FID from 139.45 to 242.10 relative to disabling ControlNet alone.

Visual conditioning. Our full model and the variant without satellite orthophoto conditioning differ only slightly across the metrics in Table 3. With satellite orthophoto conditioning, our model has lower FID (78.73 versus 81.43) and higher I_B (0.6120 versus 0.6066), together with higher I_s and mIoU and lower D_s. Its RMSE and both LPIPS scores are slightly higher, while PSNR and pixel accuracy are lower by 0.118 dB and 0.0023, respectively. The two configurations also produce visually similar results in Fig. 13(c,e).

Without a satellite orthophoto, panoramas generated from different depth maps under the same city-based text prompt may have similar visual styles, as the shared prompt provides no appearance cues specific to each location. Fig. 14 illustrates visual conditioning under a fixed city-based text prompt for Vienna. Across columns, the same satellite orthophoto produces similar visual styles under different depth conditions. Within each column, different satellite orthophotos produce different façade textures while preserving the geometry specified by the shared depth map.

5.5. Generalization Analysis on Unseen Cities

We evaluate generalization on 50 samples from each of Canberra, Johannesburg, Mexico City, and Moscow, giving 200 samples from cities excluded from training. We apply the model trained on the other 14 cities using the same inference settings, with virtual depth maps, satellite orthophotos, and city-based text prompts as inputs. No additional training is performed.

Table 4 shows that our method has the best perceptual and semantic scores and the highest I_s. Other methods perform better on D_s, RMSE, and PSNR. Its FID is 96.39, compared with 164.56 for the best baseline. In Fig. 15, our model produces clearer façade textures and follows

TABLE **4.** Quantitative comparison on unseen cities. We evaluate all four methods on 200 samples from Canberra, Johannesburg, Mexico City, and Moscow.

Method	Pixel level		Structural level		Perceptual level			Semantic level		
	RMSE↓	PSNR↑	I_s ↑	D_s ↓	P_{squeeze} ↓	P_{alex} ↓	FID↓	I_B ↑	Acc.↑	mIoU↑
Sat2Density++	**55.397**	**13.485**	0.0496	0.1626	0.5218	0.6774	197.94	0.2965	0.5465	0.1181
Sat3DGen	55.562	13.460	0.0515	**0.1482**	0.4945	0.6561	164.56	0.3472	0.6612	0.1402
CS2S	61.993	12.491	0.0414	0.1572	0.5200	0.7073	166.32	0.2378	0.5945	0.1216
Ours	64.752	11.957	**0.0564**	0.1492	**0.4442**	**0.6334**	**96.39**	**0.4105**	**0.7968**	**0.1656**

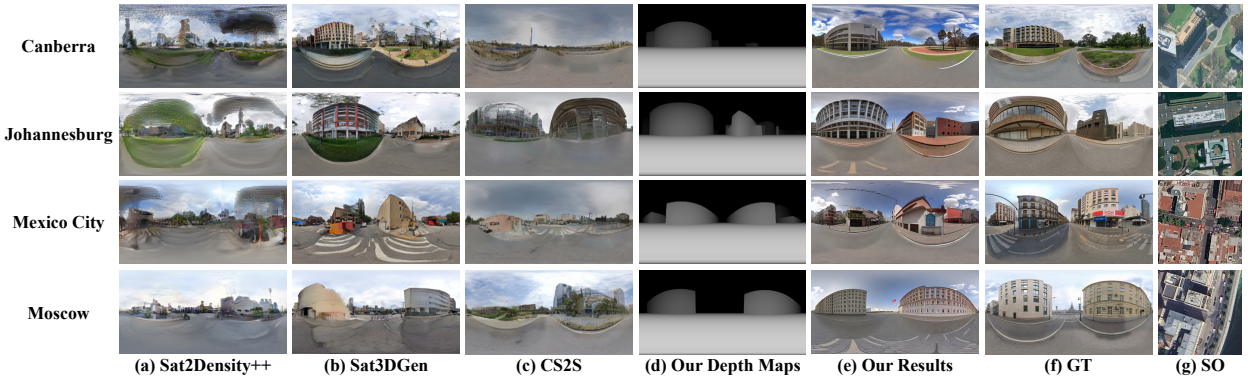


FIG. **15.** Qualitative comparison on unseen cities. Results from all four methods are shown with virtual depth maps, ground-truth panoramas, and satellite orthophoto (SO) inputs for Canberra, Johannesburg, Mexico City, and Moscow.

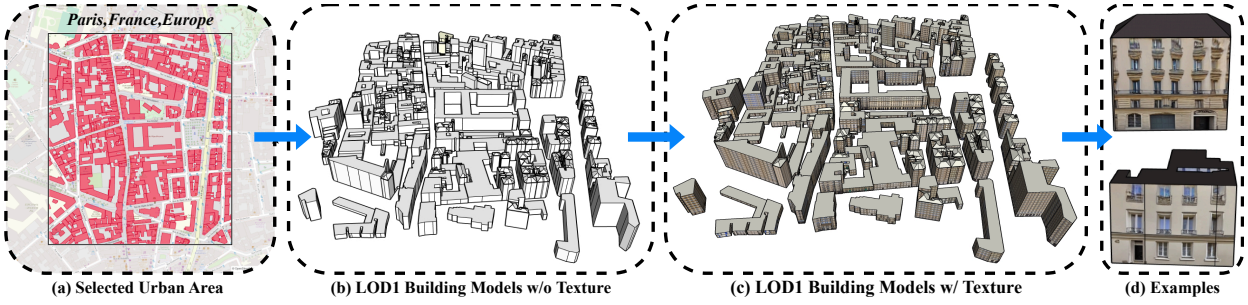


FIG. **16.** LOD-1 model construction and texture mapping in Paris. The figure shows the selected area, untextured and textured LOD-1 models, and two close-up façade examples.

the depth-specified layouts more closely than the baselines across the four unseen cities. These results show that appearance and geometric conditioning also work in unseen cities.

5.6. LOD-1 Model Construction and Texture Mapping

We apply the panorama synthesis and texture-mapping workflow to an urban area in Paris, placing sampling points at 50 m intervals along the OSM road network. At each point, we generate a virtual street-view depth map using the geometric projection described in Section 3.2 and synthesize the corresponding panorama with the trained model.

We rectify the synthesized panoramas and manually extract the visible façade regions as planar textures. We then use SketchUp PlaceMaker to construct the LOD-1 building models and map the extracted textures onto the corresponding building surfaces. Fig. 16 shows the untextured and textured models for the selected area, with two close-up examples of the mapped façades.

5.7. Limitation

Our method captures the overall visual style of building façades while following the geometry specified by virtual depth maps. However, it does not consistently reproduce fine façade details. Windows may differ in number, shape,

and arrangement from those in the reference panoramas (Fig. 12). Future work will explore façade-level visual references and explicit constraints on window layouts to guide local texture refinement.

6. Conclusion

We present a framework for texturing LOD-1 building models through street-view panorama synthesis. Virtual street-view depth maps derived from OSM building geometry constrain scene layout through ControlNet. City-based textual conditioning provides shared streetscape cues, while satellite orthophoto visual conditioning supplies local appearance information through IP-Adapter. We train ControlNet and selected IP-Adapter components in two stages with the diffusion backbone frozen. We also construct OmniCityTex, with 29,881 paired samples from 18 cities, for training and evaluation. The framework provides flexible control over façade appearance through city prompts and satellite imagery while preserving the building layout specified by the depth maps. Future work will explore façade-level visual references and explicit layout constraints to improve these details.

Acknowledgements

This study was supported in part by the National Natural Science Foundation of China (Project No.42230102, U25A20772) and the Natural Science Foundation of Sichuan Province under Grant 2026NSFSCZY0054.

References

Büyükdemircioğlu, M. and S. Oude Elberink. 2025. Generating photo-realistic urban digital twins with occlusion-aware texture mapping. *International Journal of Applied Earth Observation and Geoinformation*, **145**, 105 002

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-8 (6)**, 679–698

Chen, J., and Coauthors. 2025. China's national 3D mapping program (3DRGLM): Overall architecture and key technological issues. *Acta Geodaetica et Cartographica Sinica*, **54 (4)**, 636–649

Cheng, J., X. Huang, G. Chen, T. Wang, J. Wang, X. Tan, A. Jiang, and X. Zhang. 2026. FACTS: Training-free zero-shot diffusion framework for facade texture restoration in 3D urban models. *Advanced Engineering Informatics*, **71**, 104 385

Cohen, N. Z., O. Nir, and A. Shamir. 2025. Conditional balance: Improving multi-conditioning trade-offs in image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2641–2650

Google 2026. Street View Tiles. URL https://developers.google.com/maps/documentation/tile/streetview, last updated 10 September 2026; accessed 15 September 2026, Map Tiles API documentation, Google for Developers

Haklay, M. and P. Weber. 2008. OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, **7 (4)**, 12–18

He, H., J. Yu, P. Cheng, Y. Wang, Y. Zhu, T. Lin, and G. Dai. 2022. Automatic, multiview, coplanar extraction for CityGML building model texture mapping. *Remote Sensing*, **14 (1)**, 50

Heusel, M., H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, Vol. 30, 6626–6637

Ho, J. and T. Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*

Iandola, F. N., S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and less than 0.5 MB model size. *arXiv preprint arXiv:1602.07360*

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

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, Vol. 25, 1097–1105

Lee, J. J. and B. Benes. 2025. Top2Ground: A height-aware dual conditioning diffusion model for robust aerial-to-ground view generation. *arXiv preprint arXiv:2511.08258*, 2511.08258

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

Lin, T. J., W. Wang, Y. Shi, A. Perincherry, A. Vora, and H. Li. 2025. Geometry-guided cross-view diffusion for one-to-many cross-view image synthesis. *International Conference on 3D Vision*, 866–881

Ling, X. and R. Qin. 2023. Large-scale and efficient texture mapping algorithm via loopy belief propagation. *IEEE Transactions on Geoscience and Remote Sensing*, **61**, 1–11

Liu, X., X. Liu, Y. Zhang, W. Sun, C. Wang, T. Li, and D. Yue. 2026. All-in-one texture restoration of building façade with generative diffusion prior. *ISPRS Journal of Photogrammetry and Remote Sensing*, **239**, 379–398

Liu, Z., L. Zhai, J. Yin, X. Liu, S. Zhang, D. Wang, A. Rajabifard, and Y. Chen. 2025. TRSP: Texture reconstruction algorithm driven by prior knowledge of ground object types. *ISPRS Journal of Photogrammetry and Remote Sensing*, **223**, 221–243

Lu, X., Z. Li, Z. Cui, M. R. Oswald, M. Pollefeys, and R. Qin. 2020. Geometry-aware satellite-to-ground image synthesis for urban areas. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 859–867

Maxar 2024. Precision3D Telco Suite. Accessed 14 September 2026, Product data sheet, MXR-DS-Telco 10/24

OGC 2021. OGC City Geography Markup Language (CityGML) Part 1: Conceptual model standard. OGC Standard 20-010, Open Geospatial Consortium. Version 3.0.0

OpenAI 2026. GPT Image 2 model. URL https://developers.openai.com/api/docs/models/gpt-image-2, accessed 29 July 2026

Park, J., I.-B. Jeon, S.-E. Yoon, and W. Woo. 2021. Instant panoramic texture mapping with semantic object matching for large-scale urban scene reproduction. *IEEE Transactions on Visualization and Computer Graphics*, **27 (5)**, 2746–2756

Qian, M., B. Tan, Q. Wang, X. Zheng, H. Xiong, G.-S. Xia, Y. Shen, and N. Xue. 2026a. Seeing through satellite images at street views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **48 (5)**, 5692–5709

Qian, M., and Coauthors. 2026b. Sat3DGen: Comprehensive street-level 3D scene generation from single satellite image. *International Conference on Learning Representations*

Radford, A., and Coauthors. 2021. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Vol. 139, 8748–8763

Regmi, K. and A. Borji. 2018. Cross-view image synthesis using conditional GANs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3501–3510

Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10 684–10 695

Shang, Q., H. Hu, H. Yu, B. Xu, L. Wang, and Q. Zhu. 2024. Semantic image translation for repairing the texture defects of building models. *IEEE Transactions on Geoscience and Remote Sensing*, **62**, 1–20

Shi, Y., D. Campbell, X. Yu, and H. Li. 2022. Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44 (12)**, 10 009–10 022

Song, J., C. Meng, and S. Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*

Song, Z., X. Ze, J. Lu, and Y. Shi. 2023. Learning dense flow field for highly-accurate cross-view camera localization. *Advances in Neural Information Processing Systems*, Vol. 36, 70 612–70 625

Tang, S., Y. Chen, T. Yu, Y. Li, L. Xie, W. Wang, and R. Guo. 2026. Optimized 3D building mapping and reconstruction via cross-view collaboration in densely built-up areas. *Photogrammetric Engineering & Remote Sensing*, **92 (1)**, 49–63

Tang, W., W. Li, X. Liang, O. Wysocki, F. Biljecki, C. Holst, and B. Jutzi. 2025. Texture2LoD3: Enabling LoD3 building reconstruction with panoramic images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*

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

Tumanyan, N., O. Bar-Tal, S. Bagon, and T. Dekel. 2022. Splicing ViT features for semantic appearance transfer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10 748–10 757

Wu, S., H. Tang, X.-Y. Jing, H. Zhao, J. Qian, N. Sebe, and Y. Yan. 2023. Cross-view panorama image synthesis. *IEEE Transactions on Multimedia*, **25**, 3546–3559

Xiong, W., H. Zhang, B. Peng, Z. Hu, Y. Wu, J. Guo, and H. Huang. 2023. TwinTex: Geometry-aware texture generation for abstracted 3D architectural models. *ACM Transactions on Graphics*, **42 (6)**, 1–14

Xu, N., H. Liu, and R. Qin. 2026. Cross-view localization and synthesis: Data sets, challenges, and opportunities. *Photogrammetric Engineering & Remote Sensing*, published online 27 August 2026

Xu, N. and R. Qin. 2024. Geospecific view generation: Geometry-context-aware high-resolution ground view inference from satellite views. *European Conference on Computer Vision*, Springer, 349–366

Xu, N. and R. Qin. 2025. Satellite to GroundScape: Large-scale consistent ground view generation from satellite views. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6068–6077

Xu, N., R. Qin, D. Huang, and F. Remondino. 2024. Multi-tiling neural radiance field (NeRF)—geometric assessment on large-scale aerial datasets. *The Photogrammetric Record*, **39 (188)**, 718–740

Ye, H., J. Zhang, S. Liu, X. Han, and W. Yang. 2023. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*

Yu, F., and Coauthors. 2026. From orbit to ground: Generative city photogrammetry from extreme off-nadir satellite images. *CVPR Findings*

Ze, X., Z. Song, Q. Wang, J. Lu, and Y. Shi. 2025. Controllable satellite-to-street-view synthesis with precise pose alignment and zero-shot environmental control. *International Conference on Learning Representations*

Ze, X., B. Zhu, Z. Song, J. Lu, and Y. Shi. 2026. SatDreamer360: Multiview-consistent generation of ground-level scenes from satellite imagery. *International Conference on Learning Representations*

Zhang, L., A. Rao, and M. Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847

Zhang, M., J. Wu, Y. Liu, J. Zhang, and G. Li. 2022. GIS-based procedural modeling in 3d urban design. *ISPRS International Journal of Geo-Information*, **11 (10)**, 531

Zhang, R., P. Isola, A. A. Efros, E. Shechtman, and O. Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595

Zhang, X., and Coauthors. 2024. A 3D urban scene reconstruction enhancement approach based on adaptive viewpoint selection of panoramic videos. *The Photogrammetric Record*, **39 (185)**, 7–35

Zhou, B., H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. 2017. Scene parsing through ADE20K dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 633–641

Zhu, M., J. Yan, J. Zhong, A. Elfadaly, and L. Dai. 2026. A comparative review of representative techniques in image-based 3D dense reconstruction. *Photogrammetric Engineering & Remote Sensing*, **92 (4)**, 351–362