SynUrbanSAT: Controllable Satellite Data Synthesis for Urban Semantic 3D Reconstruction

Anonymous authors
Paper under double-blind review

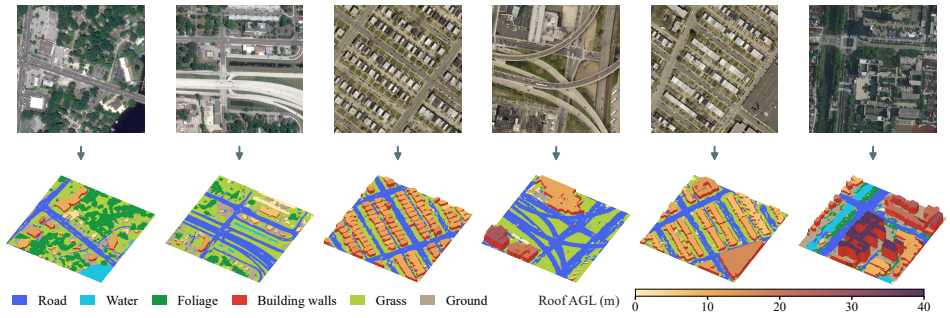


Figure 1: Showcases of urban semantic 3D reconstruction with monocular satellite imagery, enabled by **SynUrbanSAT**, a controllable and scalable satellite data synthesis pipeline.

Abstract

Urban semantic 3D reconstruction from monocular satellite imagery requires geographically diverse data with aligned semantic and height supervision, yet acquiring precise annotations at scale remains costly. Existing synthesis methods struggle to jointly provide real-world urban layouts, realistic appearance, and spatially aligned supervision. To address this gap, we propose SynUrbanSAT, a scalable and controllable pipeline that combines OpenStreetMap (OSM)-based 3D scene construction with diffusion-based satellite image synthesis. A heterogeneous dual-condition adapter (HDCA) integrates categorical semantics and continuous metric heights to guide the generation of realistic satellite imagery. The pipeline supports controllable data augmentation by varying building-height distributions, vegetation configurations, viewpoints, and appearance. It enables scalable synthesis of geographically diverse data across cities with a single trained generator, while offering the option to train a dedicated generator on limited target-city imagery for local data augmentation. To support a broad range of downstream urban perception and reconstruction tasks, we construct SynUrbanSAT-20, comprising 20,097 samples across 20 cities with synthetic RGB images and aligned semantic labels, height maps, polygon maps, point clouds, and 3D meshes. Experiments demonstrate that our image generation method improves visual fidelity and the utility of synthetic data for downstream learning. The complete pipeline further outperforms the evaluated synthetic training sources in urban semantic 3D reconstruction.

1 Introduction

In remote sensing, urban semantic 3D reconstruction from monocular satellite imagery requires both urban semantics and above-ground geometry, supporting applications such as city-scale physics simulations (Tominaga et al., 2008), urban planning (Biljecki et al., 2015), and autonomous driving (Dosovitskiy et al., 2017). However, overhead satellite imagery provides limited height cues, particularly in near-nadir views, while variations in acquisition conditions further complicate height estimation (Weir et al., 2019). Besides, robust urban understanding across regions is challenged due to substantial geographic variation in urban layouts, architectural styles, vegetation types, and ground surface appearances (Wang et al., 2021; Xia et al., 2023). Therefore, learning to recover semantic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and geometric information requires large-scale, geographically diverse training data with spatially aligned RGB imagery and supervision (Cong et al., 2022).

Although satellite imagery is abundant, acquiring precise semantic and height supervision at scale remains costly, requiring labor-intensive annotation and airborne LiDAR surveys (Xiong et al., 2023b; Toker et al., 2024). Measurement errors and misalignment between RGB imagery and height maps further compromise supervision quality (Xiong et al., 2023a; Bosch et al., 2019; Rottensteiner et al., 2014). Existing datasets also cover limited regions and provide different subsets of semantic and geometric supervision (Bosch et al., 2019; Xiong et al., 2023a; Wang et al., 2021; Weir et al., 2019; Gerke, 2015). Public map data such as OpenStreetMap (OSM) can reduce acquisition costs, but their annotations may be incomplete, inaccurate, or misaligned with satellite imagery (Haklay & Weber, 2008; Haklay, 2010). These limitations motivate synthetic data generation as a scalable alternative.

Existing synthetic approaches provide scalable, aligned supervision and broadly follow two paradigms: rendering-based approaches and diffusion-based approaches. Rendering-based approaches procedurally construct 3D urban scenes in modeling software, then jointly render aligned RGB images, semantic annotations, and height maps (Song et al., 2024a). While ensuring precise supervision, procedural generation rules and conventional rendering limit geographic fidelity and visual realism. In contrast, diffusion-based approaches generate paired RGB images and semantic annotations (Toker et al., 2024) or translate real-world layouts and semantic annotations into realistic satellite images (Sastry et al., 2024). Further, task-specific RGB references and feedback can improve controllability and transferability (Dong et al., 2025; Yang et al., 2026). However, these approaches primarily focus on 2D semantic supervision and typically lack height maps aligned with the generated imagery (Toker et al., 2024; Dong et al., 2025; Yang et al., 2026). This motivates us to explore: can we synthesize realistic satellite imagery grounded in real-world urban layouts while retaining explicit control over 3D geometry and consistency with aligned supervision?

We introduce **SynUrbanSAT**, a controllable pipeline that converts OSM urban layouts into synthetic satellite RGB images with aligned semantic annotations and metric-height maps for urban semantic 3D reconstruction. It combines precise supervision from 3D scene rendering with diffusion-based realistic RGB synthesis conditioned on semantic and height maps. The pipeline can generate large-scale remote sensing datasets by varying urban layouts, building-height distributions, vegetation configurations, and viewpoints. It can further be extended to support data augmentation by varying generation prompts, and finer-grained semantic understanding by adding new object categories. Building on this pipeline, we construct **SynUrbanSAT-20**, comprising 20,097 samples across 20 diverse cities. Beyond synthetic RGB images, semantic annotations, and metric height maps, the dataset provides aligned polygon maps, point clouds, and 3D meshes. These complementary types of supervision can support a broader range of downstream tasks training. Experimental results demonstrate improvements in both the visual quality of synthesized imagery and downstream task performance. The complete pipeline further outperforms the evaluated synthesis baselines in urban semantic 3D reconstruction. Our main contributions are:

- We introduce **SynUrbanSAT**, a controllable remote sensing data synthesis pipeline that converts real-world urban layouts and configurable 3D scenes into satellite imagery conditioned on semantic and geometric annotations. By varying building-height distributions, vegetation configurations, viewpoints, and generation prompts, this pipeline also enables scalable and controllable construction of diverse, richly annotated datasets.
- We develop **HDCA**, a heterogeneous dual-condition adapter, to adapt pretrained ControlNet to satellite image synthesis with joint semantic and metric-height conditioning. HDCA combines boundary-aware semantic encoding, Fourier-based height encoding, and adaptive fusion to produce ControlNet-compatible features.
- We construct **SynUrbanSAT-20**, comprising 20,097 samples across 20 cities with synthetic RGB images, aligned semantic and height annotations, polygon maps, point clouds, and 3D meshes to support diverse urban perception and reconstruction tasks.

2 Related Work

Real-world datasets. Learning urban semantics and geometry requires geographically diverse imagery with aligned annotations, yet existing datasets address different aspects of this require-

ment. LoveDA provides a 2D segmentation benchmark capturing urban–rural variation across three Chinese cities (Wang et al., 2021), while OpenEarthMap provides land-cover annotations worldwide (Xia et al., 2023) and SpaceNet 2 offers building footprints across four cities (Van Etten et al., 2018). Their annotations primarily support land-cover understanding and building extraction, leaving geometric supervision to other resources. US3D provides semantic labels, elevation, and point clouds for two U.S. cities (Bosch et al., 2019), while GAMUS pairs RGB imagery with nDSM and semantic annotations across five U.S. cities (Xiong et al., 2023a). ISPRS Potsdam and Vaihingen benchmarks provide semantic and height information for two German cities (Gerke, 2015). However, scaling such supervision to larger urban areas and new cities remains costly.

Rendering-based synthesis. Rendering-based approaches address the annotation bottleneck by deriving supervision from explicit 3D scenes (Khose et al., 2024; AlaaEldin et al., 2026). Procedural modeling enables systematic variation in urban layouts and scene assets (Kong et al., 2020). SyntheWorld combines this approach with diffusion-generated textures for land-cover mapping and building change detection (Song et al., 2024b), while SynRS3D additionally provides precise height supervision for semantic 3D understanding (Song et al., 2024a). Other approaches expand the sources and forms of geometric data (Chen et al., 2022; Li et al., 2023). THE obtains height supervision from game environments (Xiong et al., 2023b), SMARS simulates photogrammetric acquisition to produce orthophotos and DSMs (Fuentes Reyes et al., 2023), and UrbanScene3D provides real and synthetic scenes for multi-view reconstruction (Lin et al., 2022). To better reflect real-world urban layouts, geotypical synthesis further incorporates OSM-derived street layouts (Song et al., 2025). Although these approaches improve scene diversity and supervision availability, rule-based layouts may not fully capture the spatial layouts of real cities, while rendered appearance remains constrained by available assets, materials, and rendering pipelines. Together, these structural and visual gaps limit the fidelity of synthetic imagery to real-world scenarios.

Diffusion-based synthesis. Diffusion models offer a complementary way to improve visual realism (Ho et al., 2020; Rombach et al., 2022; Khanna et al., 2024). For controlled generation, SatSynth jointly synthesizes image–mask pairs (Toker et al., 2024), while EarthSynth introduces category control and data selection for multiple remote sensing tasks (Pan et al., 2025). Beyond semantic layouts, CRS-Diff combines text, metadata, and image-derived conditions to broaden generation control (Tang et al., 2024). To strengthen downstream utility, TISynth incorporates reference imagery for appearance transfer (Dong et al., 2025), and TODSynth guides sampling with segmentation feedback (Yang et al., 2026). Geographic grounding provides another direction: GeoSynth combines map-based layouts with location context (Sastry et al., 2024), while OSMGen and VectorSynth exploit vector geometry and semantic attributes for finer spatial control (Ziashahabi et al., 2025; Cher et al., 2026). These advances improve visual realism and spatial control, but scaling diffusion-based synthesis to diverse cities with aligned 3D supervision remains challenging. To address this gap, **SynUrbanSAT** draws on the complementary strengths of rendering-based and diffusion-based approaches to synthesize satellite RGB images from OSM urban layouts. This enables large-scale synthesis across real-world urban layouts, combining high-fidelity appearance with spatially aligned supervision.

3 Methodology

3.1 Overview

We develop a controllable urban satellite imagery synthesis pipeline, **SynUrbanSAT**, comprising three stages. As illustrated in Figure 2, we first construct 3D urban scenes from OSM data through procedural modeling in Blender, deriving semantic annotations, height and polygon maps, point clouds, and 3D meshes. Second, we train a diffusion-based generator jointly conditioned on semantic and metric-height maps to synthesize realistic satellite RGB imagery. Finally, we apply the trained generator to synthesize satellite imagery, producing a large-scale dataset with real-world urban layouts, realistic appearance, and spatially aligned supervision.

3.2 Stage 1: OSM-based 3D scene construction

Real-world urban layouts. As an open collaborative platform, OSM (Haklay & Weber, 2008) provides global data on building footprints, roads, water bodies, vegetation, and land cover. However,

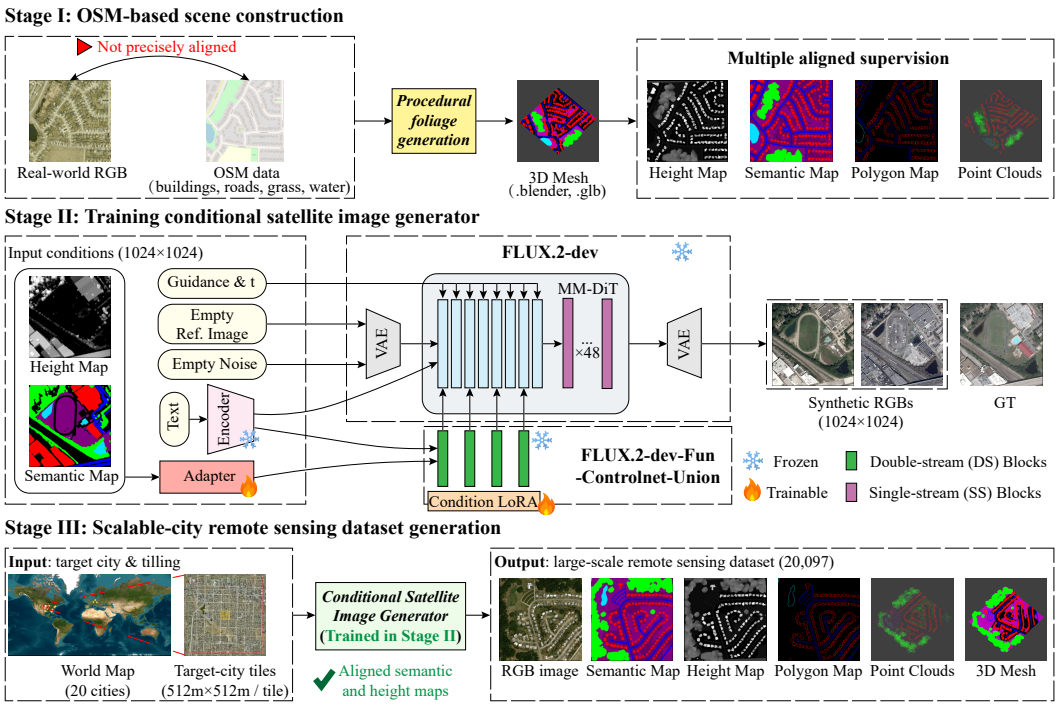


Figure 2: Overview of **SynUrbanSAT**, a controllable satellite imagery synthesis pipeline that comprises three stages: (1) Generate real-world urban layout and aligned supervision based on OSM data; (2) Train a realistic satellite image generator conditioned on semantic and height maps; (3) Generate the large-scale synthetic remote sensing dataset, **SynUrbanSAT-20**, with real-world layouts and high-fidelity visual appearance, comprising 20,097 tiles from 20 diverse cities.

irregular collection and updating limit the use of OSM annotations as direct supervision. Nevertheless, OSM captures diverse real-world urban layouts, which we use to construct 3D scenes and generate spatially aligned supervision.

3D scene construction. We first construct 3D urban scenes from OSM vectors in Blender using BlenderGIS and procedural modeling, preserving the mapped geometry of buildings, roads, water bodies, and land surfaces. We then assign building heights using valid OSM height or floor-count attributes when available; otherwise, we sample heights from predefined distributions (e.g., Gaussian or long-tailed), as OSM building-height attributes are often missing or uniform. Finally, to supplement limited vegetation information in public datasets, we procedurally generate trees in open areas with controlled density, size, and placement (Figure 3(a)). Moreover, this procedural approach allows the same urban layout to be instantiated with different height distributions for buildings and different configurations of trees, enabling controllable data augmentation (Figure 3(b)).

Multi-view rendering. Height estimation from overhead imagery is challenging because near-orthographic views provide limited geometric cues (Mou & Zhu, 2018). To represent different viewing geometries, we render each constructed 3D scene from multiple viewpoints by varying camera tilt and azimuth, as illustrated in Figure 3(c). Each viewpoint yields a spatially aligned pair of semantic annotations and height maps. The 3D mesh, sampled point cloud, and nadir-view polygon map are shared across all views of the same scene.

3.3 Stage 2: Conditioned Satellite Imagery Generator

Architecture. We build the generator on pretrained FLUX.2-dev (Black Forest Labs, 2025) and FLUX.2-dev-Fun-Controlnet-Union (FunCN) (Alibaba PAI, 2026). To adapt FunCN from natural images to overhead satellite imagery, we introduce a *heterogeneous dual-condition adapter* (HDCA) and apply parameter-efficient LoRA (Hu et al., 2022) updates (Figure 2). As shown in Figure 4, HDCA uses separate encoding branches to accommodate categorical semantic labels and continuous

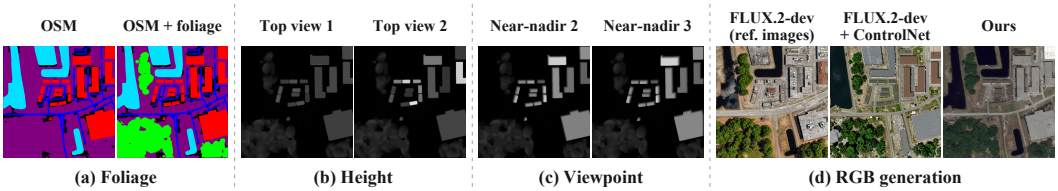


Figure 3: OSM scene controls and RGB synthesis. (a) Procedural foliage generation. (b) Building-height variations. (c) Viewpoints: 15°/135° and 25°/225° (tilt/azimuth). (d) FLUX.2-dev with reference images, FunCN, and ours with fixed inputs. Prompt details are provided in Appendix A.2.

heights: the semantic branch combines learned class embeddings with boundary-enhanced convolutions, while the height branch uses Fourier features (Tancik et al., 2020). The resulting features are fused, projected, and downsampled to match FunCN's latent space. Then, FunCN processes these features and injects the resulting residuals into FLUX.2-dev's double-stream layers 0, 2, 4, and 6. Only HDCA and the LoRA parameters on FunCN are optimized, while the pretrained FLUX.2-dev and FunCN weights remain frozen. Details are provided in Appendix A.3.

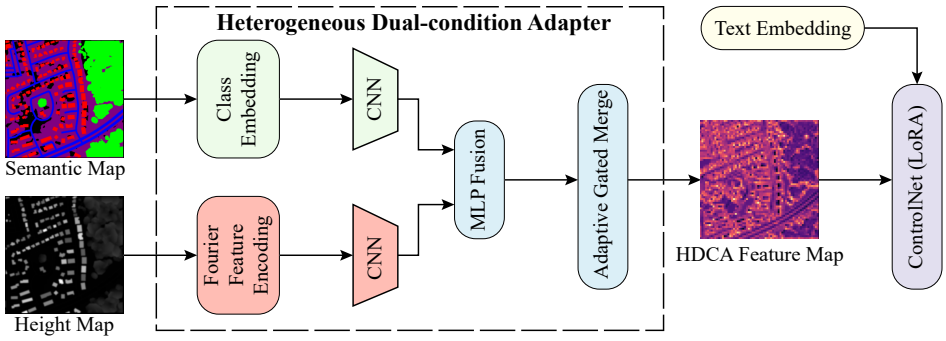


Figure 4: The architecture of HDCA, fusing categorical semantic labels and continuous heights.

Training dataset. To adapt the generator to real-world satellite imagery, we select 125 spatially non-overlapping RGB–semantic–height triplets of size 2048×2048 from US3D (Bosch et al., 2019). We manually relabel the semantic annotations to correct original annotation errors and divide each triplet into four 1024×1024 crops to match the generator's input resolution, yielding **US3D-Enhanced** with 500 aligned samples. These are split by source group into $400/48/52$ training/validation/test samples. The semantic annotations before and after manual relabeling are compared in Appendix A.4.

Training configuration. The generator is trained at 1024×1024 resolution with a weighted MSE loss on predicted flow velocities. We use a fixed satellite prompt and a guidance scale of 3.5, without data augmentation. AdamW is used with learning rates of 10^{-3} for HDCA and 10^{-4} for LoRA, weight decay of 0.01, and gradient clipping at 1.0. Learning rates are linearly warmed up over 400 updates from 1% of their base values, followed by cosine decay to 10^{-7}. Training runs for 300 epochs, with the selected checkpoint based on validation loss and qualitative inspection of generated images. Additional training details are provided in Appendix A.5.

3.4 Stage 3: Satellite Imagery Synthesis Pipeline

Synthesis settings. Our pipeline supports two synthesis settings. For cross-city synthesis, a generator is trained once and applied to urban layouts from other cities to construct geographically diverse data. For city-specific synthesis, a separate generator can be trained on a small amount of real imagery from a target city to synthesize satellite images that match its local appearance. Both settings enable scalable and cost-effective augmentation of real-world datasets.

SynUrbanSAT-20 dataset. Following the cross-city synthesis setting, we use the generator trained on US3D-Enhanced to construct **SynUrbanSAT-20**, comprising 20,097 tiles from 20 cities. Each tile pairs generated RGB imagery with aligned semantic annotations, height and polygon maps, point

clouds, and 3D meshes. Figure 5 shows representative samples, while Figure 6(a) summarizes city coverage and class distribution. In addition, samples from **SynUrbanSAT-20** achieve the lowest DINO-FID/DINO-KID among other synthetic sources relative to US3D-Enhanced, indicating higher visual fidelity (Figure 6(b)). Geographic extents and class statistics are provided in Appendix A.6.

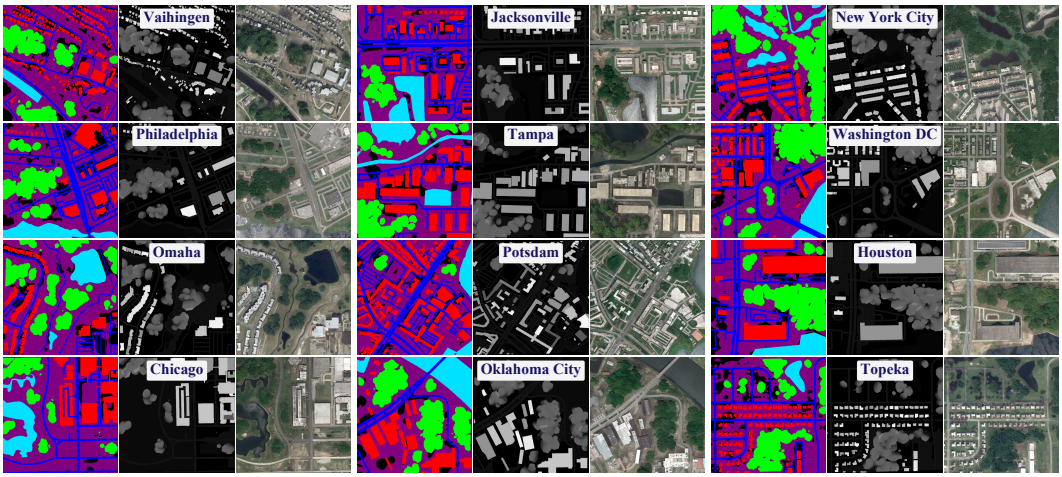
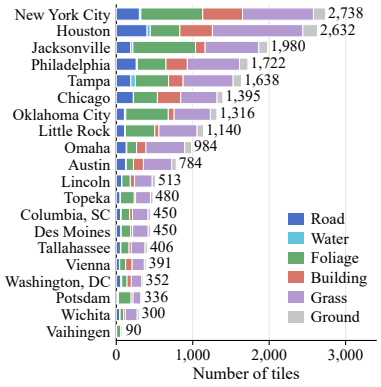
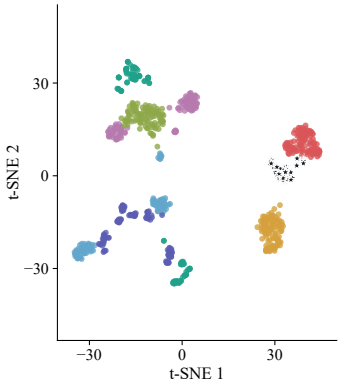


Figure 5: Samples (semantic annotations, height maps, synthetic RGBs) from **SynUrbanSAT-20**.



(a) City coverage (20 cities / 20,097 tiles)

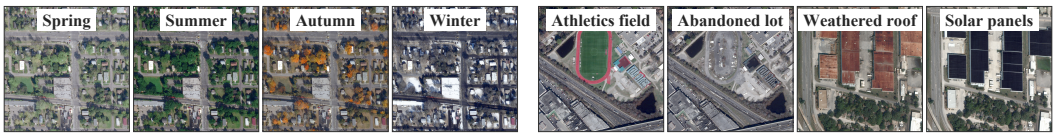


Source	DINO-FID ↓	DINO-KID ↓ ($\times 10^3$)
★ US3D-Enhanced (reference)	—	—
SynRS3D	0.876	1.511
SMARS	0.952	1.677
SyntheWorld	0.697	1.141
SyntCities	1.092	1.848
TopAir	1.099	1.635
FLUX.2 + FunCN	0.483	0.683
Ours	**0.338**	**0.394**

(b) DINOv2 feature space and fidelity

Figure 6: Dataset coverage and fidelity. (a) City tile counts and semantic class distribution. (b) Joint DINOv2 t-SNE and DINO-FID/DINO-KID (↓), DINO-KID is scaled by 10^3.

Extended controllability. Beyond these synthesis settings, our pipeline supports data augmentation to capture appearance variation across acquisition time and conditions. By varying text prompts, for example, it generates imagery with different seasonal characteristics and local appearances (Figure 7). The detailed prompts are provided in Appendix A.2.



(a) Season **(b) Local appearance**

Figure 7: Prompt-controlled appearance. (a) Seasonal appearance variations. (b) Local appearance variations: athletics field/abandoned lot (left) and weathered/solar roofs (right).

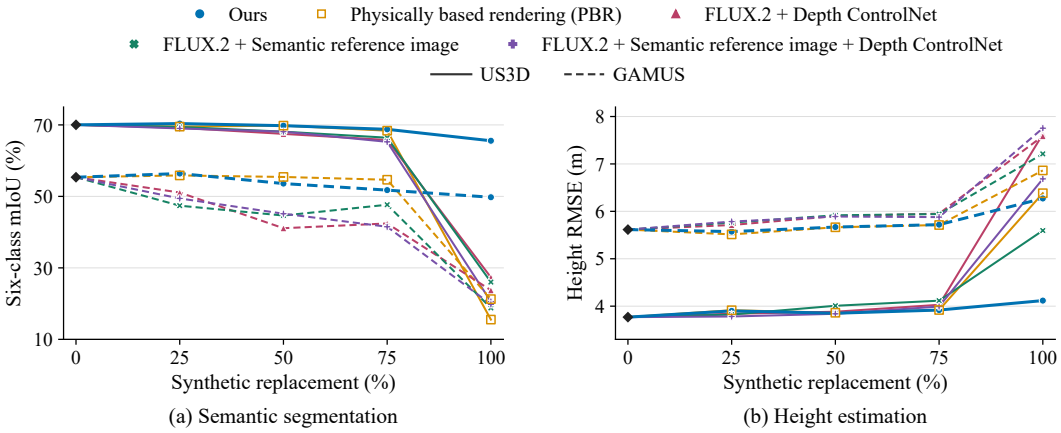
4 Experiments

4.1 Experimental Objectives and Setup

To distinguish the contribution of RGB generation from the overall effectiveness of the pipeline, we evaluate both the RGB generation stage and the complete synthesis pipeline. Section 4.2 compares our RGB generation method with other synthesis approaches in terms of generation quality and downstream task performance. Section 4.3 compares the complete pipeline with other synthetic training sources for urban semantic 3D reconstruction. For downstream evaluation, separate DINOv2 (ViT-L/14) models with DPT decoders predict semantic labels and heights, evaluated using mIoU and RMSE, respectively. Model and data processing details are provided in Appendix A.7.

4.2 Generation Quality and Downstream Utility

Comparison with Blender rendering and Flux.2 generations. We hold US3D-Enhanced layouts and supervision fixed and progressively replace real-world RGB with synthetic images across 400 training samples. As shown in Figure 8, our generated RGB maintains comparatively stable downstream performance on test data from US3D (in-domain) and GAMUS (out-of-domain) as the replacement ratio increases, with gradual changes in segmentation mIoU and height RMSE. In contrast, the PBR (physically based rendering) baseline, which uses RGB images rendered directly in Blender, and the FLUX.2-based baselines exhibit pronounced degradation when real training imagery is fully replaced. These results suggest that our realistic RGB synthesis helps narrow the sim-to-real gap, preserving the utility of synthetic data for downstream tasks on real-world scenarios.



(a) Semantic segmentation (b) Height estimation

Figure 8: Downstream performance across synthetic replacement ratios on US3D (in-domain) and GAMUS (out-of-domain): (a) Semantic segmentation mIoU; (b) Height estimation RMSE.

Comparison with diffusion-based texturing. To compare our full-image generation with diffusion-based texturing, we use SynRS3D (Song et al., 2024a), whose ground sampling distance (GSD), semantic categories, and class distribution are comparable to those of our data. We use the semantic annotations and height maps from 400 SynRS3D layouts both to condition our RGB generator and as labels for downstream training. With layouts, labels, and training settings held fixed, we vary only the RGB source, as illustrated in Figure 9. The results of downstream performance are shown in Table 1. Our generated RGB improves semantic segmentation on four of five datasets and reduces height estimation RMSE on US3D and GAMUS. These results support the advantage of our full-image generation over diffusion-based texturing.

4.3 Urban Semantic 3D Reconstruction

We evaluate the complete pipeline through urban semantic 3D reconstruction from a single satellite image, using building volume IoU to jointly assess footprint and height accuracy. Height extrusion is restricted to buildings because extruding vegetation or elevated roads would incorrectly fill the open spaces beneath them with solid volumes. Avoiding these artifacts is important for applications

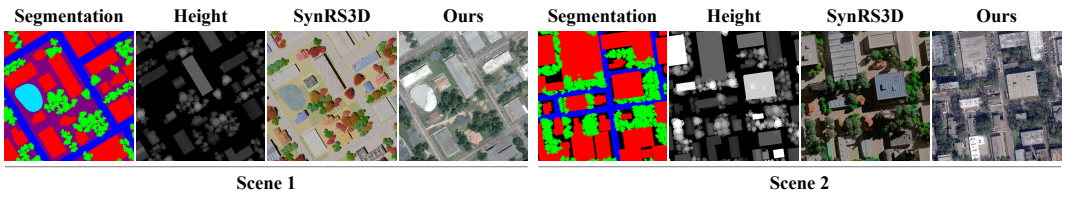


Figure 9: Qualitative comparison on two SynRS3D scenes with shared semantic and height conditions. Experimental results demonstrate that our method improves both the visual quality of synthesized imagery and downstream task performance compared with diffusion-based texturing.

Table 1: Downstream performance when training on rendered versus generated RGB images with fixed SynRS3D layouts and labels. RBW stands for road, building, and water. Bold: best.

	US3D		GAMUS		LoveDA	Potsdam	SpaceNet 2
RGB	RBW mIoU (%) ↑	Height RMSE (m) ↓	RBW mIoU (%) ↑	Height RMSE (m) ↓	RBW mIoU (%) ↑	Building IoU (%) ↑	Building IoU (%) ↑
Rendered	61.65	7.15	66.19	11.26	**48.08**	69.17	61.24
Ours	**64.74**	**5.27**	**67.80**	**9.09**	46.66	**73.28**	**64.30**

that rely on physically meaningful urban geometry, such as seismic risk assessment, CFD simulation, and solar access analysis (Biljecki et al., 2015).

We compare our synthetic training data with SynRS3D, SMARS (Fuentes Reyes et al., 2023), SyntCities (Fuentes Reyes et al., 2022), and TopAir (AlaaEldin et al., 2026) on US3D and GAMUS, using 400 training images per source. We consider two configurations of our data: **OSM-Fixed-400** uses a shared training set generated from layouts of Jacksonville and Omaha, the two US3D cities, for evaluation on both datasets; **OSM-Target-City-400** uses layouts selected according to the target dataset, drawing from Jacksonville and Omaha for US3D, and New York City, Philadelphia, and Washington, D.C. for GAMUS.

Table 2: Results of urban 3D reconstruction. Bold: best; underline: second best.

	Building Volume IoU (%) ↑					
Test dataset	SynRS3D	SMARS	SyntCities	TopAir	Ours	
					OSM-Fixed-400	OSM-Target-City-400
US3D	26.10	24.76	8.15	<u>30.44</u>	**37.09**	**37.09**
GAMUS	31.91	36.23	12.72	41.40	<u>43.71</u>	**44.11**

As shown in Table 2, OSM-Fixed-400 achieves building volume IoUs of 37.09/43.71% on US3D/GAMUS, exceeding the strongest external baseline, TopAir, by 6.65/2.31 pp. These gains extend to the out-of-domain GAMUS dataset, demonstrating the transferability of our synthetic training data. Our pipeline further enables imagery generation from the layouts of specific target cities, allowing the synthetic data to better match local urban structure and spatial distributions. Using target-city layouts, OSM-Target-City-400 increases the GAMUS score to 44.11%, supporting the benefit of tailoring synthetic training data to the target domain. Figure 10 compares semantic 3D reconstruction results across other synthetic training sources.

4.4 Ablation Study

To assess the effectiveness of generator adaptation, we compare our adapted generator (HDCA + LoRA) with three unadapted FLUX.2-dev configurations: conditioning on a Blender-rendered PBR RGB reference image, a single ControlNet conditioned on height maps, and two summed ControlNet

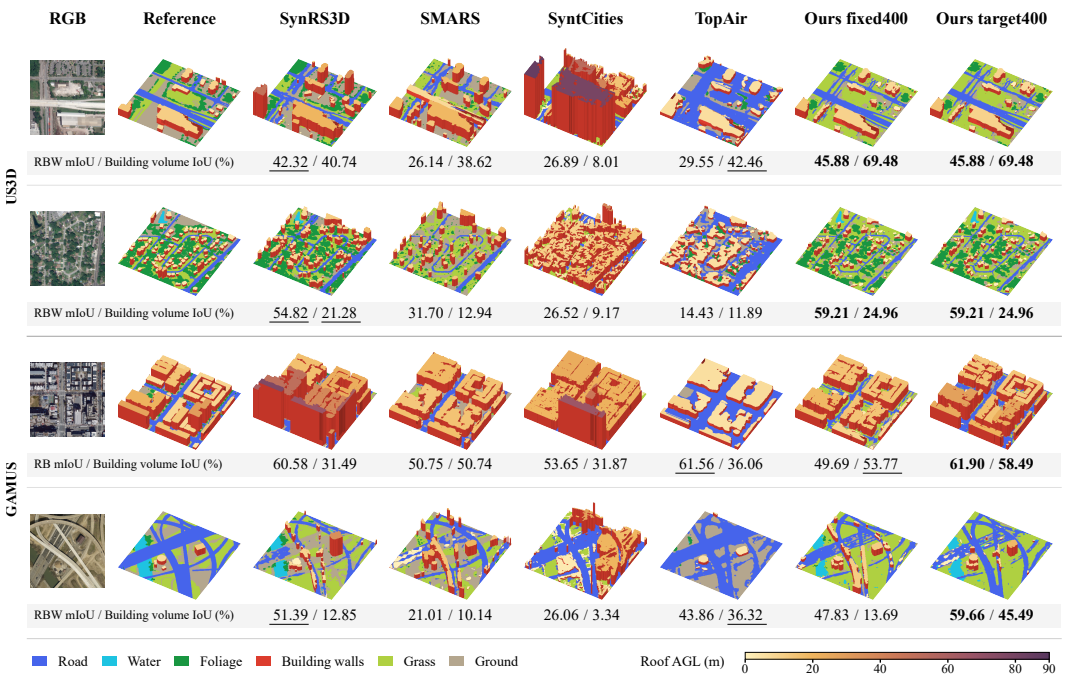


Figure 10: Comparison of urban semantic 3D reconstruction results on US3D (top) and GAMUS (bottom).

branches conditioned separately on semantic annotations and height maps. All configurations use the same OSM-Fixed-400 layouts and annotations under a fixed downstream training budget.

Table 3: Ablation study results. Bold: best.

Configuration	mIoU (%) ↑			Height RMSE (m) ↓		Building volume IoU (%) ↑	
	US3D	GAMUS	LoveDA	US3D	GAMUS	US3D	GAMUS
FLUX.2 (PBR reference)	20.20	23.27	19.83	**6.91**	7.44	6.87	19.86
FLUX.2 + depth ControlNet	18.35	17.53	18.27	7.19	7.78	17.75	26.32
FLUX.2 + S+H ControlNets	22.25	14.90	19.29	7.16	7.78	20.40	28.37
Ours (HDCA + LoRA)	**49.41**	**36.06**	**30.87**	6.95	**7.23**	**37.09**	**43.71**

As shown in Table 3, our adapted generator achieves the highest segmentation mIoU across all three datasets and the highest building volume IoU on US3D and GAMUS. These results highlight the importance of satellite-specific adaptation in enhancing the ability of generation foundation model to generate more reliable training data for downstream urban perception and reconstruction.

5 Conclusion

We presented **SynUrbanSAT**, a scalable and controllable pipeline that converts OSM urban layouts into realistic satellite imagery with aligned supervision through 3D scene rendering and diffusion-based RGB generation. HDCA integrates semantic and height conditions to guide imagery synthesis, while configurable scenes enable the construction of diverse, richly annotated datasets. The resulting **SynUrbanSAT-20** comprises 20,097 samples across 20 cities with complementary 2D and 3D annotations for urban perception and reconstruction. Experiments demonstrate improvements in generation quality and downstream performance, with the complete pipeline outperforming the evaluated synthetic training sources in urban semantic 3D reconstruction.

AI use statement

In this work, we used generative AI tools (GPT-6) to assist language polishing. The authors personally reviewed and revised the final version of the manuscript and take responsibility for all its content. Our proposed method uses generation model (FLUX.2-dev) to generate synthetic satellite RGB imagery. This data generation process is the core of the research methodology, and the RGB images in the proposed **SynUrbanSAT-20** dataset are synthetic data.

Ethics statement

All authors have reviewed the ICLR 2027 Code of Ethics and conducted this research in accordance with its principles. Our use of publicly available models and datasets complies with their respective licenses and terms of use.

Reproducibility statement

We will release the code for our dataset construction pipeline, providing a scalable and controllable approach (**SynUrbanSAT**) to generating synthetic remote sensing datasets for the community. We will also release the trained generator (checkpoint, LoRA and HDCA parameter) and the synthetic dataset (**SynUrbanSAT-20**) introduced in this work in accordance with applicable licenses. The training configurations and procedures, data splits, and input–output processing are detailed in the paper and Appendices. Besides, the core codes of downstream tasks and main experiments will also be released.

References

Yara AlaaEldin, Francesca Odone, and Mohamed Maher Ata. A study on synthetic-to-real generalization of depth estimation and semantic segmentation joint networks on aerial images. *Complex & Intelligent Systems*, 12:201, 2026. doi: 10.1007/s40747-026-02292-6. URL https://link.springer.com/article/10.1007/s40747-026-02292-6.

Alibaba PAI. FLUX.2-dev-Fun-Controlnet-Union. Official model card, 2026. URL https://huggingface.co/alibaba-pai/FLUX.2-dev-Fun-Controlnet-Union. Accessed September 21, 2026.

Filip Biljecki, Jantien Stoter, Hugo Ledoux, Sisi Zlatanova, and Arzu Çöltekin. Applications of 3D city models: State of the art review. *ISPRS International Journal of Geo-Information*, 4(4):2842–2889, 2015. doi: 10.3390/ijgi4042842. URL https://research.tudelft.nl/en/publications/applications-of-3d-city-models-state-of-the-art-review/.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.01401.

Black Forest Labs. FLUX.2 [dev]. Official model card, 2025. URL https://huggingface.co/black-forest-labs/FLUX.2-dev. Accessed September 21, 2026.

Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D. Hager, and Myron Brown. Semantic stereo for incidental satellite images. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 1524–1532, 2019. doi: 10.1109/WACV.2019.00167. URL https://arxiv.org/abs/1811.08739.

Meida Chen, Qingyong Hu, Zifan Yu, Hugues Thomas, Andrew Feng, Yu Hou, Kyle McCullough, Fengbo Ren, and Lucio Soibelman. STPLS3D: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. In *British Machine Vision Conference*, 2022. URL https://arxiv.org/abs/2203.09065.

Daniel Cher, Brian Wei, Srikumar Sastry, and Nathan Jacobs. VectorSynth: Fine-grained satellite image synthesis with structured semantics. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7019–7029, 2026. doi: 10.1109/WACV61042.2026.00678.

URL https://openaccess.thecvf.com/content/WACV2026/papers/Cher_VectorSynth_Fine-Grained_Satellite_Image_Synthesis_with_Structured_Semantics_WACV_2026_paper.pdf.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems*, volume 35, pp. 197–211, 2022. doi: 10.52202/068431-0015. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/01c561df365429f33fcd7a7faa44c985-Abstract-Conference.html.

Runmin Dong, Shuai Yuan, Litong Feng, Jinxiao Zhang, Weijia Li, Mengxuan Chen, Bin Luo, Wayne Zhang, and Haohuan Fu. Transferable image synthesis for remote sensing semantic segmentation via joint reference-semantic fusion. *Information Fusion*, 127:103839, 2025. doi: 10.1016/j.inffus.2025.103839. URL https://github.com/dongrunmin/TISynth.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 1–16, 2017. URL https://proceedings.mlr.press/v78/dosovitskiy17a.html.

Mario Fuentes Reyes, Pablo d'Angelo, and Friedrich Fraundorfer. SyntCities. Zenodo, version 1, 2022. URL https://zenodo.org/records/6967325.

Mario Fuentes Reyes, Yuxing Xie, Xiangtian Yuan, Pablo d'Angelo, Franz Kurz, Daniele Cerra, and Jiaojiao Tian. A 2d/3d multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205:74–97, 2023. doi: 10.1016/j.isprsjprs.2023.09.013. URL https://elib.dlr.de/198879/.

Markus Gerke. Report for 2015, scientific initiative: ISPRS benchmark on urban object detection and 3D building reconstruction. Technical report, International Society for Photogrammetry and Remote Sensing, 2015. URL https://www.isprs.org/documents/si/SI-2015/ISPRS-SI2015-Gerke_report_WGIII4_2015.pdf.

Mordechai Haklay. How good is volunteered geographical information? a comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37(4):682–703, 2010. doi: 10.1068/b35097. URL https://www.homepages.ucl.ac.uk/~ucfamha/OSM%20data%20analysis%20070808_web.pdf.

Mordechai Haklay and Patrick Weber. OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008. doi: 10.1109/MPRV.2008.80. URL https://veiv.cs.ucl.ac.uk/people/students/56-publications/232-openstreetmap-user-generated-street-maps.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. DiffusionSat: A generative foundation model

for satellite imagery. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/16c3c941409d0581286eff49b180930f-Abstract-Conference.html.

Sahil Santosh Khose, Anisha Pal, Aayushi Agarwal, Deepanshi, Judy Hoffman, and Prithvijit Chattopadhyay. SkyScenes: A synthetic dataset for aerial scene understanding. In *European Conference on Computer Vision*, pp. 19–35, 2024. URL https://eccv.ecva.net/virtual/2024/poster/1729.

Fanjie Kong, Bohao Huang, Kyle Bradbury, and Jordan Malof. The Synthinel-1 dataset: A collection of high resolution synthetic overhead imagery for building segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1814–1823, 2020. doi: 10.1109/WACV45572.2020.9093339. URL https://openaccess.thecvf.com/content_WACV_2020/html/Kong_The_Synthinel-1_dataset_a_collection_of_high_resolution_synthetic_overhead_WACV_2020_paper.html.

Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. MatrixCity: A large-scale city dataset for city-scale neural rendering and beyond. In *IEEE/CVF International Conference on Computer Vision*, pp. 3205–3215, 2023. doi: 10.1109/ICCV51070.2023.00297. URL https://openaccess.thecvf.com/content/ICCV2023/html/Li_MatrixCity_A_Large-scale_City_Dataset_for_City-scale_Neural_Rendering_and_ICCV_2023_paper.html.

Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: the UrbanScene3D dataset. In *European Conference on Computer Vision*, pp. 93–109, 2022. doi: 10.1007/978-3-031-20074-8_6. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/3218_ECCV_2022_paper.php.

Lichao Mou and Xiao Xiang Zhu. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network, 2018. URL https://arxiv.org/abs/1802.10249.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Jiancheng Pan, Shiye Lei, Yuqian Fu, Jiahao Li, Yanxing Liu, Yuze Sun, Xiao He, Long Peng, Xiaomeng Huang, and Bo Zhao. EarthSynth: Generating informative earth observation with diffusion models, 2025. URL https://arxiv.org/abs/2505.12108.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ranftl_Vision_Transformers_for_Dense_Prediction_ICCV_2021_paper.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Franz Rottensteiner, Gunho Sohn, Markus Gerke, Jan Dirk Wegner, Uwe Breitkopf, and Jaewook Jung. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93:256–271, 2014. doi: 10.1016/j.isprsjprs.2013.10.004.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, and Nathan Jacobs. GeoSynth: Contextually-aware high-resolution satellite image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 460–470, 2024. doi: 10.1109/CVPRW63382.2024.00051. URL https://arxiv.org/abs/2404.06637.

Jian Song, Hongruixuan Chen, Weihao Xuan, Junshi Xia, and Naoto Yokoya. SynRS3D: A synthetic dataset for global 3d semantic understanding from monocular remote sensing imagery. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024a. doi: 10.52202/079017-3727. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d4f3da38b491b44d40a0d5a5b37134ba-Abstract-Datasets_and_Benchmarks_Track.html.

Jian Song, Hongruixuan Chen, and Naoto Yokoya. SyntheWorld: A large-scale synthetic dataset for land cover mapping and building change detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8287–8296, 2024b. URL https://openaccess.thecvf.com/content/WACV2024/html/Song_SyntheWorld_A_Large-Scale_Synthetic_Dataset_for_Land_Cover_Mapping_and_WACV_2024_paper.html.

Shuang Song, Yang Tang, and Rongjun Qin. Synthetic data matters: Retraining with geo-typical synthetic labels for building detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025. doi: 10.1109/TGRS.2025.3593864. URL https://arxiv.org/abs/2507.16657.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/55053683268957697aa39fba6f231c68-Abstract.html.

Datao Tang, Xiangyong Cao, Xingsong Hou, Zhongyuan Jiang, Junmin Liu, and Deyu Meng. CRS-Diff: Controllable remote sensing image generation with diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3453414. URL https://arxiv.org/abs/2403.11614.

Aysim Toker, Marvin Eisenberger, Daniel Cremers, and Laura Leal-Taixé. SatSynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27695–27705, 2024. doi: 10.1109/CVPR52733.2024.02615. URL https://openaccess.thecvf.com/content/CVPR2024/html/Toker_SatSynth_Augmenting_Image-Mask_Pairs_through_Diffusion_Models_for_Aerial_Semantic_CVPR_2024_paper.html.

Yoshihide Tominaga, Akashi Mochida, Ryuichiro Yoshie, Hiroto Kataoka, Tsuyoshi Nozu, Masaru Yoshikawa, and Taichi Shirasawa. AIJ guidelines for practical applications of CFD to pedestrian wind environment around buildings. *Journal of Wind Engineering and Industrial Aerodynamics*, 96(10–11):1749–1761, 2008. doi: 10.1016/j.jweia.2008.02.058. URL https://www.niit.ac.jp/kotozukuri/resource_db/labo-26/achievement/.

Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. SpaceNet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. doi: 10.48550/arXiv.1807.01232. URL https://arxiv.org/abs/1807.01232.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/4e732ced3463d06de0ca9a15b6153677-Abstract-round2.html.

Nicholas Weir, David Lindenbaum, Alexei Bastidas, Adam Van Etten, Sean McPherson, Jacob Shermeyer, Varun Kumar, and Hanlin Tang. SpaceNet MVOI: A multi-view overhead imagery dataset. In *IEEE/CVF International Conference on Computer Vision*, pp. 992–1001, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Weir_SpaceNet_MVOI_A_Multi-View_Overhead_Imagery_Dataset_ICCV_2019_paper.html.

Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. OpenEarthMap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6254–6264, 2023. URL https://openaccess.thecvf.com/content/WACV2023/html/Xia_OpenEarthMap_A_Benchmark_Dataset_for_Global_High-Resolution_Land_Cover_Mapping_WACV_2023_paper.html.

Zhitong Xiong, Sining Chen, Yi Wang, Lichao Mou, and Xiao Xiang Zhu. GAMUS: A geometry-aware multi-modal semantic segmentation benchmark for remote sensing data, 2023a. URL https://arxiv.org/abs/2305.14914.

Zhitong Xiong, Wei Huang, Jingtao Hu, and Xiao Xiang Zhu. THE benchmark: Transferable representation learning for monocular height estimation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023b. doi: 10.1109/TGRS.2023.3311764. URL https://arxiv.org/abs/2112.14985.

Yunkai Yang, Yudong Zhang, Kunquan Zhang, Jinxiao Zhang, Xinying Chen, Haohuan Fu, and Runmin Dong. Task-oriented data synthesis and control-rectify sampling for remote sensing semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42147–42157, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Yang_Task-Oriented_Data_Synthesis_and_Control-Rectify_Sampling_for_Remote_Sensing_Semantic_CVPR_2026_paper.pdf.

Amir Ziashahabi, Narges Ghasemi, Sajjad Shahabi, John Krumm, Salman Avestimehr, and Cyrus Shahabi. OSMGen: Highly controllable satellite image synthesis using OpenStreetMap data. In *NeurIPS UrbanAI Workshop*, 2025. URL https://arxiv.org/abs/2511.00345.

A Appendix

A.1 Procedural Height Assignment and Control

A.1.1 Building Height Assignment

In Stage 1, we preserve OSM building footprints and extrude each polygon to its assigned height. The implementation also supports attribute-based assignment when individual OSM features retain valid attributes: a finite height tag takes precedence, followed by the number of floors converted at 3 m per floor, with procedural sampling used when neither attribute is available. When the input contains only merged footprint geometry, each resulting polygon receives a sampled height.

Let $C_{a,b}(x) = \min\{b, \max\{a, x\}\}$ denote clipping to $[a, b]$. The default procedural height sampler uses a clipped log-normal distribution:

$$h_i^{\text{sample}} = C_{a,b}(\exp(\mu + \sigma\epsilon_i)), \qquad \epsilon_i \sim \mathcal{N}(0, 1), \tag{1}$$

The default bounds are $a = 3$ m and $b = 30$ m. With $m = (a + b)/2$, the distribution parameters are

$$\mu = \log m, \qquad \sigma = \max\left\{\frac{\log b - \mu}{1.96}, 0.15\right\}. \tag{2}$$

Samples outside the specified interval are clipped rather than resampled. The implementation also supports a constant height m, a uniform distribution $\mathcal{U}(a, b)$, and a clipped Gaussian mixture. The mixture selects the lower and upper components with probabilities 0.7 and 0.3, respectively. Their means are $a + (b-a)/3$ and $b - (b-a)/3$, and their standard deviations are $\max\{1, 0.08(b-a)\}$ and $\max\{1, 0.15(b-a)\}$ meters. The default building-height sampling seed is 42. Features are sorted by OSM identifier and geometry bounds, or by bounds when identifiers are unavailable, to make sampling reproducible for a fixed input.

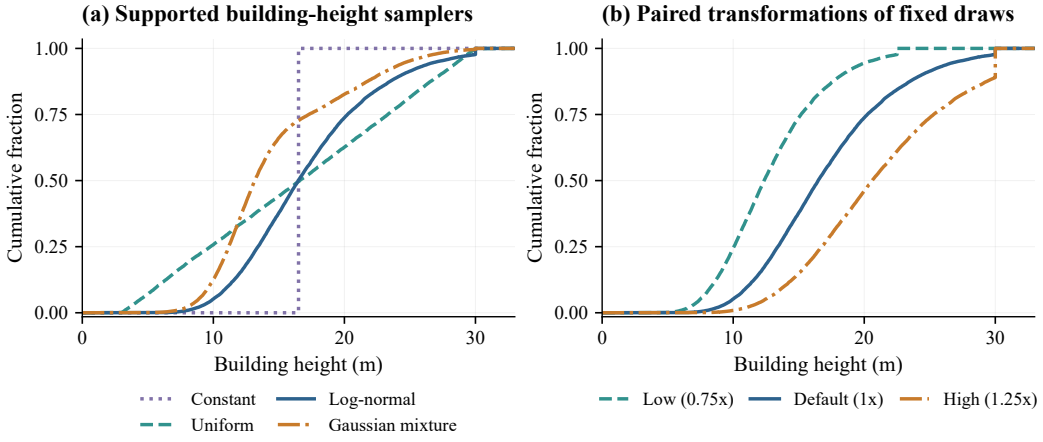


Figure 11: Procedural building-height controls. (a) Empirical cumulative distributions of 20,000 draws per sampler, using bounds $[3, 30]$ m and seed 42. Constant sampling places all mass at 16.5 m. (b) Paired transformations of the same log-normal draws. The curves illustrate the samplers and do not represent dataset height statistics.

A.1.2 Paired Height Variations at a Fixed Layout

Buildings. Given an initial building height h_i^0, a separate control produces a height variant:

$$h_i^{(v)} = \max\{0.5, C_{a_v,b_v}(s_v h_i^0)\}, \tag{3}$$

Here, $s_v > 0$ and the clipping bounds are configurable. The transformation also applies when h_i^0 is obtained from an OSM height tag or floor count. Holding the source geometry, sampling seed, and initial heights fixed isolates the effect of height variation. Figure 11(b) shows the low, default,

and high settings $(s_v, a_v, b_v) = (0.75, 3, 25)$, $(1, 3, 30)$, and $(1.25, 4, 30)$, respectively, with heights measured in meters.

Trees. Procedural foliage generation uses Geometry Nodes to instantiate trees and filter intersections with buildings, roads, and water, followed by the configured canopy-overlap filter. In the default mode, tree heights vary through asset geometry and node scaling, subject to a final height cap. An optional absolute-height mode supports paired height variations by assigning each retained tree a reproducible quantile $u_j \in (0, 1)$. The quantile depends on the seed, layout identity, original horizontal position, and asset template, but not on the requested height range. Changing the range therefore preserves the correspondence between trees and quantiles. Uniform sampling assigns

$$t_j^{(v)} = \ell_v + u_j(r_v - \ell_v), \tag{4}$$

where $[\ell_v, r_v]$ specifies the desired height interval. Constant-height control assigns the interval midpoint. The absolute-height mode supports these two sampling options.

After applying the tree's object transform to its mesh, we modify only the vertex heights:

$$z' = z_{\min} + \frac{t_j^{(v)}}{z_{\max} - z_{\min}}(z - z_{\min}). \tag{5}$$

Horizontal coordinates and object translations remain unchanged. Collision and canopy-overlap filtering precede height assignment and are not repeated afterward, so the retained tree instances are identical across variants. The realized mesh height is checked against the target with a tolerance of 10^{-4} m. Figure 12 illustrates this paired mapping using recorded tree identities and positions.

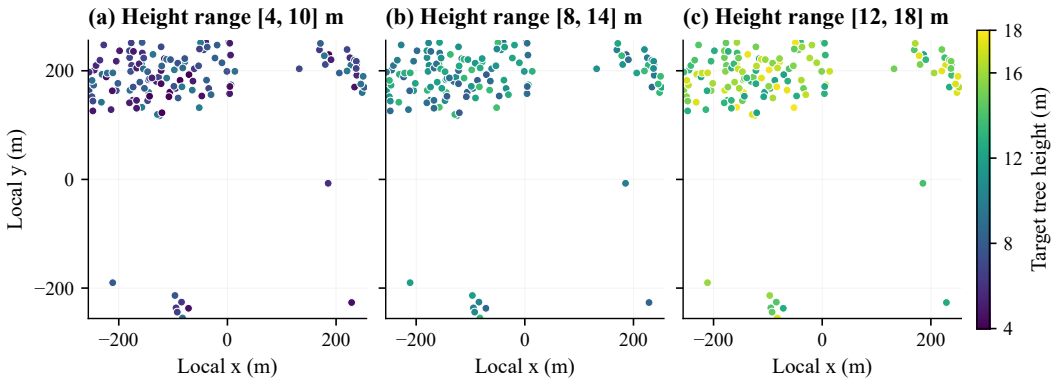


Figure 12: Paired tree-height control at fixed horizontal locations. Each point represents one of 138 retained trees from a stored scene. The same recorded quantiles are mapped to $[4, 10]$, $[8, 14]$, and $[12, 18]$ m using Equation 4; locations, instance count, and color scale are shared. The last range agrees with the recorded target heights. Panels show deterministic height assignments rather than new scene renders; marker size does not represent crown size.

Controllable variation. By sampling different building heights and vegetation configurations, including placement and height, our pipeline can generate diverse scene variants from a fixed underlying urban layout for data augmentation. These variations may improve the robustness of downstream models by exposing them to a broader range of urban configurations. Due to space limitations, we do not experimentally evaluate these augmentation benefits in this work and leave their systematic assessment to future work.

A.2 Prompts for Satellite Image Synthesis

Table 4: Semantic and height conditions and generation prompts used in Figures 3 and 7.

Figure and conditions	Prompt
Fig. 3(d) OSM + foliage in Fig. 3(a); top view 1 in Fig. 3(b)	**RGB generation (all three methods).** Aerial top-down satellite view of an American urban area, Google Earth style, photorealistic satellite imagery, natural daylight. Use the supplied layout maps only to determine geometry and spatial arrangement. Detailed weathered gray concrete and muted brown roofs with HVAC units; asphalt roads with crisp lane markings; reflective dark water; natural green tree canopies, muted lawns and bare soil. Bright natural daylight, realistic sun shadows, clear visibility, crisp details, natural color grading. Render a natural aerial photograph rather than the artificial colors of the input maps, with no labels or text.
Fig. 7(a) US3D `JAX_Tile_018_007_1`	**Spring.** Aerial top-down satellite view of American urban area, Google Earth style, 8k resolution, photorealistic satellite imagery, natural daylight, buildings: detailed rooftops with HVAC units, weathered concrete, roads: asphalt streets with lane markings and realistic texture, water: reflective surfaces with realistic ripples and color variation, bright natural daylight, realistic sun shadows, clear visibility, masterpiece, raw photo, crisp details, natural color grading, spring season, fresh pale green tree canopies, blooming trees, bright new grass, clear realistic satellite texture
Fig. 7(a) US3D `JAX_Tile_018_007_1`	**Summer.** Aerial top-down satellite view of American urban area, Google Earth style, 8k resolution, photorealistic satellite imagery, natural daylight, buildings: detailed rooftops with HVAC units, weathered concrete, roads: asphalt streets with lane markings and realistic texture, water: reflective surfaces with realistic ripples and color variation, bright natural daylight, realistic sun shadows, clear visibility, masterpiece, raw photo, crisp details, natural color grading, summer season, dense deep green tree canopies, lush saturated lawns, strong bright daylight, realistic satellite texture
Fig. 7(a) US3D `JAX_Tile_018_007_1`	**Autumn.** Aerial top-down satellite view of American urban area, Google Earth style, 8k resolution, photorealistic satellite imagery, natural daylight, buildings: detailed rooftops with HVAC units, weathered concrete, roads: asphalt streets with lane markings and realistic texture, water: reflective surfaces with realistic ripples and color variation, bright natural daylight, realistic sun shadows, clear visibility, masterpiece, raw photo, crisp details, natural color grading, autumn season, vivid orange and yellow tree canopies, scattered fallen leaves, muted dry grass, natural satellite colors
Fig. 7(a) US3D `JAX_Tile_018_007_1`	**Winter.** Aerial top-down satellite view of American urban area, Google Earth style, 8k resolution, photorealistic snowy satellite imagery, cold winter daylight, buildings: detailed rooftops with HVAC units, weathered concrete, roads: asphalt streets with lane markings and realistic texture, water: reflective surfaces with realistic ripples and color variation, ground: snow-covered lawns, parking lots, paved surfaces, and road edges, vegetation: leafless bare trees, snow-dusted branches, dormant shrubs, bright cold winter daylight, realistic snow shadows, clear visibility, masterpiece, raw photo, crisp details, natural color grading, winter season, clear snow cover on roofs, lawns and open ground, leafless trees, plowed dark roads, realistic snowy satellite texture, fresh snow on rooftops and ground, leafless bare trees
Fig. 7(b) US3D `JAX_Tile_110_023_1`	**Athletics field.** Aerial top-down satellite view of American urban area, Google Earth style, 8k resolution, photorealistic satellite imagery, natural daylight, buildings: detailed rooftops with HVAC units, weathered concrete, roads: asphalt streets with lane markings and realistic texture, water: reflective surfaces with realistic ripples and color variation, central_oval_site: the large central oval site is a school athletics ground with a clearly visible red rubber running track, crisp white lane markings, a green grass infield, school sports facilities, and small bleachers; a broad multi-lane urban roadway runs diagonally along the lower-left edge of the scene, with dark asphalt pavement, clear lane markings, and sparse vehicles, and this diagonal feature is clearly a road rather than a river, bright natural daylight, realistic sun shadows, clear visibility, masterpiece, raw photo, crisp details, natural color grading

Table 4: (continued)

Figure and conditions	Prompt
Fig. 7(b) US3D JAX_Tile_ 110_023_1	**Abandoned lot.** Aerial top-down satellite view of American urban area, Google Earth style, 8k resolution, photorealistic satellite imagery, natural daylight, buildings: detailed rooftops with HVAC units, weathered concrete, roads: asphalt streets with lane markings and realistic texture, water: reflective surfaces with realistic ripples and color variation, central_oval_site: the large central oval site is an abandoned parking lot with cracked gray asphalt, faded white parking bays, broken pavement, sparse weeds growing through cracks, scattered debris, and no active vehicles; a broad multi-lane urban roadway runs diagonally along the lower-left edge of the scene, with dark asphalt pavement, clear lane markings, and sparse vehicles, and this diagonal feature is clearly a road rather than a river, bright natural daylight, realistic sun shadows, clear visibility, masterpiece, raw photo, crisp details, natural color grading
Fig. 7(b) US3D JAX_Tile_ 173_021_2	**Weathered roof.** Aerial top-down satellite view of an American industrial warehouse district, Google Earth style, 8k resolution, photorealistic satellite imagery, natural daylight, buildings: large low-rise warehouse buildings with aged industrial roofs, muted rusty reddish-brown corrugated metal, uneven gray-brown oxidation patches and darker rain streaks, small rooftop vents, roads: asphalt streets with lane markings and realistic texture, water: reflective surfaces with realistic ripples and color variation, layout: retain the existing long rectangular warehouse footprints, roof heights, surrounding streets, parking areas, and trees; roof details stay inside the existing building footprints, bright natural daylight, realistic sun shadows, clear visibility, masterpiece, raw photo, crisp details, natural color grading
Fig. 7(b) US3D JAX_Tile_ 173_021_2	**Solar panels.** Aerial top-down satellite view of an American industrial warehouse district, Google Earth style, 8k resolution, photorealistic satellite imagery, natural daylight, buildings: large low-rise warehouse buildings with orderly dense rows of small dark navy photovoltaic panels on existing flat roofs, narrow pale maintenance aisles between rectangular panel arrays, realistic fine silver panel frames, existing rooftop vents left unobstructed, roads: asphalt streets with lane markings and realistic texture, water: reflective surfaces with realistic ripples and color variation, layout: retain the existing long rectangular warehouse footprints, roof heights, surrounding streets, parking areas, and trees; roof details stay inside the existing building footprints, bright natural daylight, realistic sun shadows, clear visibility, masterpiece, raw photo, crisp details, natural color grading

A.3 Generator Architecture and Adaptation

A.3.1 Architecture and Feature Dimensions

Condition encoding. HDCA encodes categorical semantic annotations and continuous metric heights in separate branches. Semantic maps are decoded into six class indices and mapped to learned 64-dimensional embeddings. Boundary enhancement compares each embedding with its 3×3 local average; locations where the summed absolute difference exceeds 0.1 receive an additional learned 3×3 convolutional feature. In the height branch, non-finite values are replaced with zero, and above-ground-level (AGL) heights are normalized as $u = \text{clip}(h, 0, 30)/30$. This fixed normalization is shared across tiles and affects only the generator input; the stored metric-height annotations remain unchanged. Fourier encoding concatenates u with $\sin(2\pi 2^j u)$ and $\cos(2\pi 2^j u)$ for $j = 0, \ldots, 31$, producing 65 channels.

Each branch applies three 3×3 convolutions with stride 2 and channel widths $128 \to 256 \to 512$, with group normalization and GELU after each convolution. A linear projection then produces 512-dimensional tokens. For a 1024×1024 condition map, each branch outputs a $128 \times 128 \times 512$ feature grid.

Fusion and projection. The semantic and height features are concatenated and fused by a residual MLP with layer normalization and dimensions $1024 \to 1024 \to 1024$. The fused output is split into two 512-dimensional streams, each projected to 3072 dimensions. A sigmoid gate predicts a scalar weight at each spatial location to blend the two projected streams. A learned global sigmoid

scale, initialized with a logit of -5, controls the initial adapter magnitude. Finally, 2×2 average pooling produces $64 \times 64 \times 3072$ features compatible with the conditioning input of FLUX.2-dev-Fun-Controlnet-Union (FunCN).

A.3.2 ControlNet Adaptation

FunCN processes the HDCA features and injects the resulting residuals into double-stream layers 0, 2, 4, and 6 of FLUX.2-dev. We apply LoRA to the query, key, and value projections of both the image and text streams, and to the image-stream output projection, in each of the four ControlNet attention blocks. For a frozen projection W, the adapted operation is $Wx + (\alpha/r)BAx$, with rank $r = 512$, scaling parameter $\alpha = 512$, and no dropout. We jointly train HDCA and these LoRA factors on real satellite imagery to adapt the conditioning pathway to urban structure and appearance. The pretrained FLUX.2-dev weights, remaining FunCN parameters, and VAE remain frozen.

A.4 Construction of US3D-Enhanced

We construct US3D-Enhanced from US3D imagery of Jacksonville and Omaha to train the conditional satellite image generator and evaluate RGB synthesis with fixed layouts and supervision. The dataset contains 125 source tiles of size 2048×2048. Each tile is divided into four 1024×1024 crops using identical windows for the RGB imagery, semantic annotations, and AGL height maps. This procedure yields 500 aligned RGB–semantic–height samples grouped by their 125 source tiles.

A.4.1 Semantic and Height Map Processing

The original semantic annotations contain class errors and boundaries that do not consistently align with the RGB imagery. We relabel them using six classes: road, water, foliage, building, grass, and ground. Annotation is performed in CVAT with Segment Anything Model (SAM3)-assisted segmentation, followed by manual inspection. Vegetation, particularly tree canopies and grass, remains difficult to delineate even with manual annotation. SAM3 can also select large contiguous vegetation regions, leading to local boundary misalignment. Buildings, roads, and water can generally be delineated more accurately and aligned more closely with the imagery.

We retain the original AGL height maps, including occasional spikes inherited from the source LiDAR measurements. Median filtering was examined as a way to suppress these artifacts, but its smoothing can introduce additional misalignment with RGB object boundaries. We therefore use the original AGL maps for training and evaluation. Figure 13 compares the original and enhanced semantic annotations with the corresponding RGB imagery and retained height maps.

A.4.2 Dataset Split

We partition US3D-Enhanced by source tile, assigning all four crops from a 2048×2048 tile to the same split. The training, validation, and test sets contain $400/48/52$ samples from $100/12/13$ source groups, respectively. The generator is trained on the 400 training triplets, selected using the 48 validation triplets, and evaluated for downstream utility using the 52 held-out test triplets. Grouping by source tile prevents adjacent crops from the same image from appearing in different splits.

A.5 Generator Training and Sampling

A.5.1 Flow-Matching Objective

The frozen VAE encodes each RGB image into latent tokens $\boldsymbol{z}_0$. Given Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and timestep t, we construct the noisy latent and target velocity as

$$\boldsymbol{z}_t = (1-t)\boldsymbol{z}_0 + t\epsilon, \qquad \boldsymbol{v}^* = \epsilon - \boldsymbol{z}_0. \tag{6}$$

The conditional generator predicts the flow velocity $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t; \boldsymbol{s}, \boldsymbol{h}, \boldsymbol{p})$ using the noisy latent, semantic map $\boldsymbol{s}$, height map $\boldsymbol{h}$, and prompt embedding $\boldsymbol{p}$. During training, we sample $t = \mathrm{sigmoid}(a)$ with $a \sim \mathcal{N}(0, 1)$ and minimize the weighted flow-matching objective:

$$\mathcal{L}_{\text{train}} = \frac{1}{B}\sum_{i=1}^{B} \widetilde{w}_i \,\mathrm{MSE}(\boldsymbol{v}_{\boldsymbol{\theta},i}, \boldsymbol{v}_i^*), \qquad \widetilde{w}_i = \frac{w_i}{B^{-1}\sum_{j=1}^{B} w_j}, \quad w_i = \frac{1}{t_i(1-t_i) + 10^{-3}}, \tag{7}$$

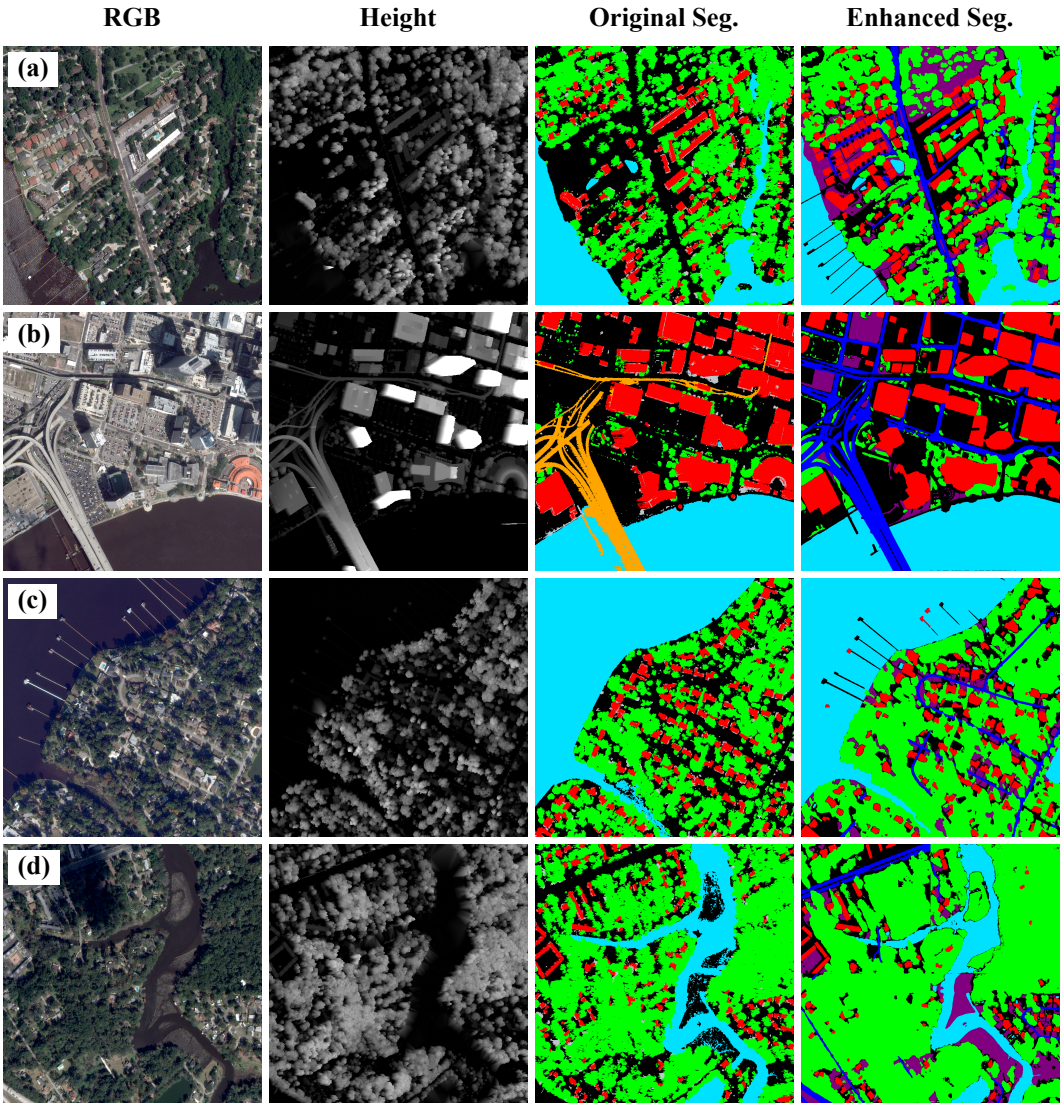


Figure 13: Four 2048×2048 source tiles from US3D-Enhanced, comparing RGB imagery, retained AGL height maps, original semantic annotations, and enhanced semantic annotations. Relabeling corrects class assignments and object boundaries, while the source heights are retained.

Here, MSE averages over latent tokens and channels, and B denotes the minibatch size. Validation uses uniformly sampled timesteps and unweighted MSE. A fixed validation seed of 640048 makes checkpoint comparisons reproducible.

A.5.2 Optimization and Sampling

The selected generator (e300 checkpoint) uses a saved embedding of the following prompt for training and default inference: "Aerial top-down satellite view of American urban area, Google Earth style, 8k resolution, photorealistic satellite imagery, natural daylight, buildings: detailed rooftops with HVAC units, weathered concrete, roads: asphalt streets with lane markings and realistic texture, vegetation: lush green trees and grass areas, ground: natural earth tones, parking lots, paved surfaces, water: reflective surfaces with realistic ripples and color variation, bright natural daylight, realistic sun shadows, clear visibility, masterpiece, raw photo, crisp details, natural color grading". Table 5 summarizes the optimization settings. Computation uses BF16, while frozen linear weights are stored in FP8 to reduce memory use.

Table 5: Generator adaptation settings for the e300 checkpoint.

Item	Setting
Resolution / precision	1024×1024 / BF16
Batch size	12; accumulation over 4 minibatches (nominal effective batch 48)
Optimizer	AdamW; weight decay 0.01; gradient clipping at 1.0
Learning rates	HDCA 10^{-3}; LoRA 10^{-4}
Schedule	400-update warmup from 1% of base rates; cosine decay to 10^{-7}
LoRA	Rank $r = 512$; scaling parameter $\alpha = 512$; dropout 0
Selected checkpoint	e300; 300 complete epochs; 2,700 optimizer updates
Augmentation / seed	None / 64

We train the generator for 300 complete epochs. With 400 training samples, each epoch contains nine optimizer updates, including the final partially accumulated batch, giving 2,700 updates in total. Among the evaluated checkpoints, e300 has the lowest fixed-validation MSE (0.83358), with qualitative inspection also assessing appearance and adherence to the semantic and height conditions.

Figure 14 shows the recorded learning curves for the training run that produces e300. Periodic fixed-validation evaluations show decreasing MSE, reaching 0.83358 at e300, with smaller improvements toward the end of training. Because training and validation use different timestep distributions and loss weighting, their numerical difference does not measure a train–validation generalization gap.

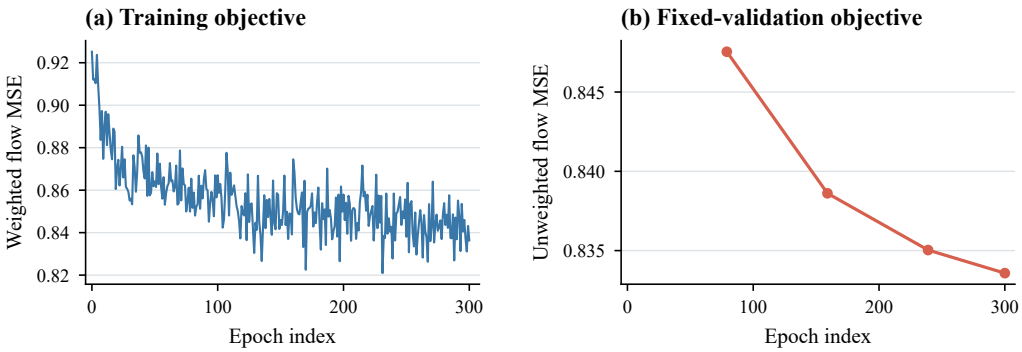


Figure 14: Learning curves for the training run that produces the e300 generator used in the main experiments. Recorded training losses and periodic fixed-validation measurements are shown without smoothing; the two objectives differ as defined in Section A.5.

At inference, sampling starts from Gaussian noise and integrates the predicted velocity from $t = 1$ to $t = 0$ using 28 uniformly spaced Euler steps and guidance scale 3.5. Per-sample seeds are derived deterministically from root seed 64. The default control strength is 1.0, as in training; the strength-2 comparison uses 2.0. Semantic and height preprocessing is identical during training and inference. Reference RGB imagery is not used as an inference condition.

A.5.3 Hyperparameter Analysis

LoRA rank. We compare LoRA ranks $r \in \{128, 256, 512\}$ with the HDCA architecture held fixed (Table 6). Figure 15 shows the training and validation curves over the first 300 recorded epochs; the table averages losses over the common interval of logged epoch indices 280–299. Increasing rank modestly reduces the mean training loss, but validation MSE does not improve monotonically. Because effective batch sizes and learning-rate schedule horizons also vary across these runs, the comparison does not isolate the effect of rank. The e300 generator used in the main experiments adopts $r = 512$, with its training settings specified in Table 5.

Table 6: LoRA-rank comparison with a fixed HDCA architecture. Losses average logged epoch indices 280–299; B is the nominal effective batch size and T the scheduled update horizon.

Rank r	Trainable (M)	B	T	Train loss	Val. MSE
128	48.77	24	4,800	0.8529	0.8433
256	88.09	24	8,000	0.8488	0.8477
512	166.73	48	4,000	0.8483	0.8457

The rank study uses a separate 400/80 training/validation split and relative-height preprocessing; its losses are not directly comparable with the e300 validation loss.

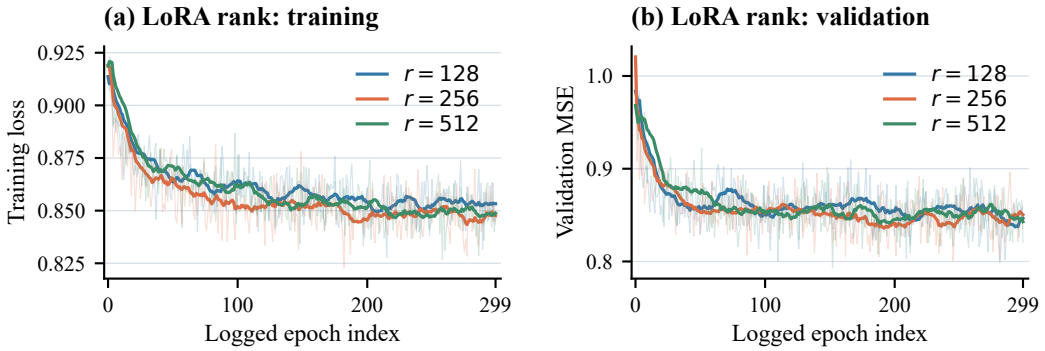


Figure 15: LoRA-rank comparison with a fixed HDCA architecture, using the runs in Table 6. Panels show (a) training loss and (b) validation MSE. Thin lines show recorded epoch-level values; solid lines show trailing 15-epoch means. All curves use the same displayed range of logged epoch indices 0–299.

A.6 SynUrbanSAT-20 Dataset Statistics

A.6.1 Tile Definition and Geographic Coverage

SynUrbanSAT-20 comprises 20,097 OSM-derived tiles from 20 cities. For each city, we specify a rectangular geographic region and partition it into non-overlapping 512×512 m cells in local UTM coordinates. The final row and column extend to cover the selected region. Each tile contains 1024×1024 pixels at a ground sampling distance of 0.5 m/pixel, giving a total nominal footprint of 5,268.3 km^2. Figure 6(a) summarizes city coverage and class composition, and Figure 16 shows the corresponding geographic extents and tile grids.

A.6.2 Class Distribution

Figure 17 reports the six-class composition of all tiles in each city's selected region. For city c, the proportion of class k is $p_{c,k} = \sum_{i \in c} n_{i,k} / \sum_{i \in c} \sum_{j=1}^{6} n_{i,j}$, where $n_{i,k}$ is the number of pixels assigned to class k in tile i. We count one semantic map per spatial tile, so multiple viewpoints do not duplicate a layout in these statistics. All 21,073,231,872 pixels in the 20,097 maps belong to the six-class palette.

Figure 18 compares all 20,097 **SynUrbanSAT-20** layouts with six annotated datasets. Row labels identify the available collections and sample counts. Class proportions are computed from pooled native-resolution pixels after taxonomy mapping, excluding unlabeled and unmapped pixels. For US3D, we use the training and validation annotations from US3D-Enhanced.

To compare datasets with different taxonomies, we map trees and forest to foliage, and grass, low vegetation, rangeland, and agriculture to grass/crops. Bare land and developed space in SynRS3D and SyntheWorld are mapped to ground. These coarse groups do not imply identical annotation conventions across datasets. SMARS does not annotate water as a separate class, as indicated by the N/A

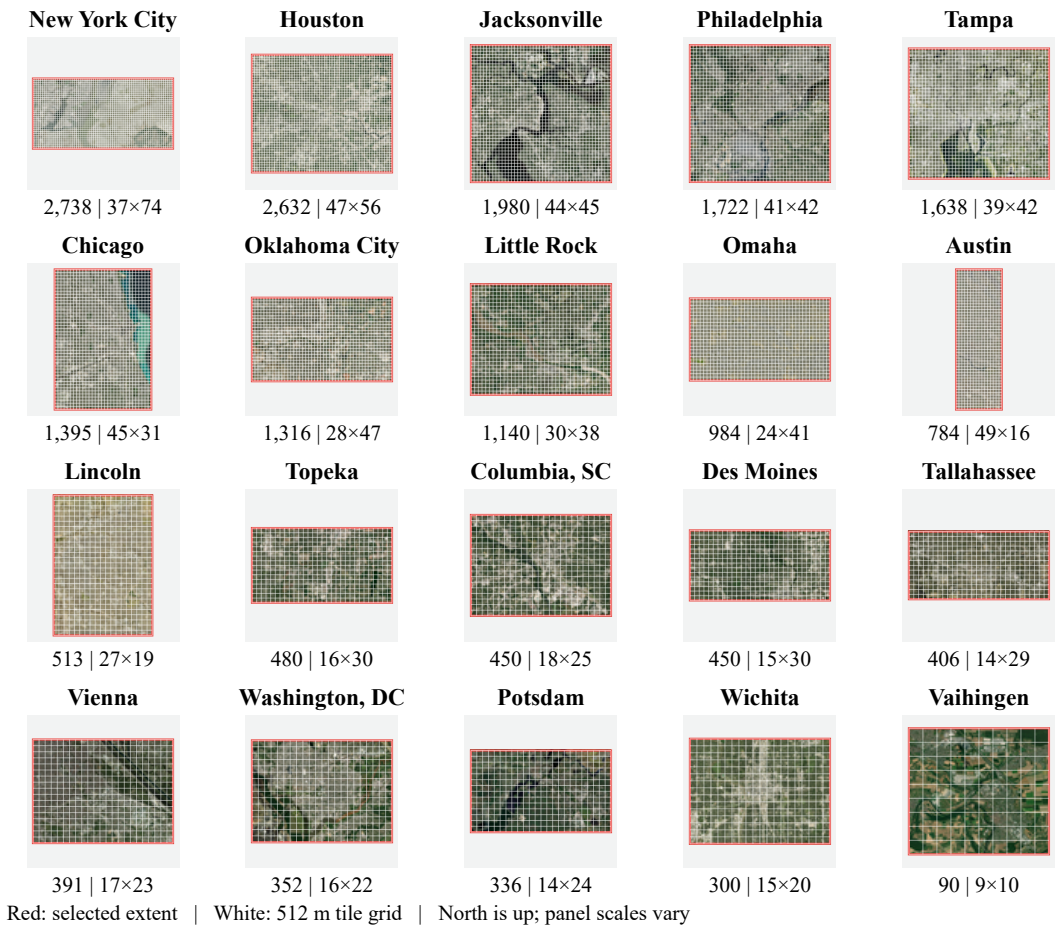


Figure 16: Geographic coverage of the 20 cities (4×5). Red frames mark grid extents; white lines indicate 512×512 m tiles. Labels show tile counts and grid dimensions (rows $\times$ columns) from the OSM tiling configuration. Each north-up panel preserves geographic aspect ratio at its own scale. Basemap: Esri World Imagery.

footnote. Grass/crops (40.1%) and foliage (26.6%) account for the largest shares of **SynUrbanSAT-20**, whereas SynRS3D has a larger building share (40.7%).

A.6.3 Visual Fidelity in DINOv2 Feature Space

We evaluate visual fidelity using frozen DINOv2 ViT-L/14 features (Oquab et al., 2024). RGB images are center-cropped to a square, bicubically resized to 518×518, and normalized with ImageNet statistics. The resulting 1024-dimensional global features are L2-normalized. The real-image reference consists of all 48 US3D-Enhanced validation images from 12 source groups disjoint from generator training.

Each synthetic source—SMARS, SyntheWorld, SynRS3D, SyntCities (Fuentes Reyes et al., 2022), TopAir, FLUX.2 with FunCN, and **SynUrbanSAT**—contributes 100 unique images, selected before feature extraction by hash ordering with seed 64. SyntCities and TopAir use the 400-image collections from the downstream comparisons, with duplicates removed by RGB hash; TopAir uses its urban nadir subset. We retain the extracted features for the five existing synthetic datasets and the real-image reference, and verify numerical agreement with the shared DINOv2 checkpoint by re-extracting eight samples. For **SynUrbanSAT**, the e300 generator with control strength 1 contributes 50 images each from Jacksonville and Omaha. The fidelity comparison therefore uses different scene collections and covers two cities for our method, while the class-distribution analysis covers all 20 cities.



Figure 17: City-level class composition across all 20,097 layouts. Panels (a) and (b) show ten cities each, in the same order as Figure 16. Each stacked bar pools native-resolution semantic pixels within one city and sums to 100%; labels give the corresponding tile counts.

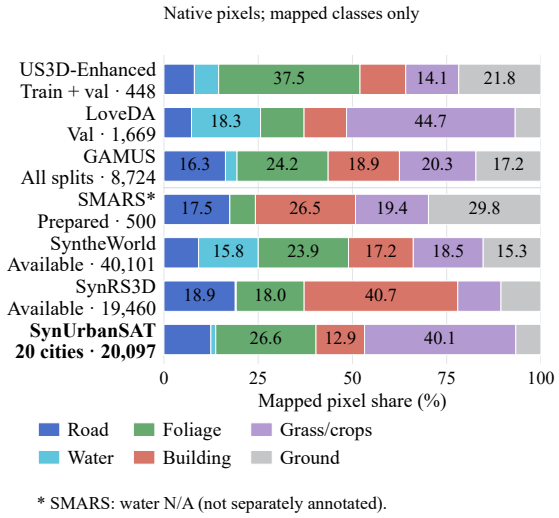


Figure 18: Class composition across real and synthetic datasets after coarse taxonomy mapping. Bars show pooled pixel percentages; row labels specify the available splits or collections and mask counts. The SMARS footnote denotes an unannotated water class.

Figure 6(b) visualizes the features using joint t-SNE after 50-dimensional PCA, with perplexity 30, 1,500 iterations, and seed 64. DINO-FID applies the FID formulation (Heusel et al., 2017) to the DINOv2 features, measuring the Fréchet distance between their empirical means and covariances. DINO-KID applies the KID formulation (Bińkowski et al., 2018) as an unbiased squared maximum mean discrepancy with kernel $k(x, y) = (1+x^\top y/1024)^3$, excluding within-set self-pairs. Both met-

rics are computed from the original L2-normalized 1024-dimensional features before PCA or t-SNE. On these fixed samples, **SynUrbanSAT** achieves the lowest distances among the seven synthetic sources: DINO-FID 0.338 and DINO-KID 0.394×10^{-3}. SyntCities scores 1.092 and 1.848×10^{-3}, and TopAir scores 1.099 and 1.635×10^{-3}. These results measure feature-space appearance similarity to the US3D validation reference, which is also used for generator checkpoint selection; they do not establish a general ranking of human-perceived realism.

A.7 Downstream Training and Evaluation

A.7.1 Model, Preprocessing, and Training

We train separate DINOv2 ViT-L/14 models (Oquab et al., 2024) with compact DPT decoders (Ranftl et al., 2021) for semantic segmentation and height estimation. RGB images are bilinearly resized to 518×518 and normalized with ImageNet statistics. The encoder produces 37×37 feature maps with 1024 channels. Each decoder combines four feature levels at a common width of 256 channels and outputs predictions at 518×518. The segmentation model predicts six class channels, while the height model predicts a single non-negative AGL map in meters. We train the decoder and the encoder's final four transformer blocks, keeping the remaining encoder parameters frozen.

Semantic annotations are resized using nearest-neighbor interpolation. Height annotations use validity-aware bilinear interpolation, excluding non-finite and negative values. Spatial resizing preserves the metric height scale, and predictions and reference maps are evaluated on the same 518×518 grid. Preprocessing and backbone adaptation are identical across training-data sources for each task.

Optimization. For comparisons with 400 training images per source, we optimize each task independently with AdamW, weight decay 10^{-4}, batch size 8, and FP32 precision. Learning rates are fixed at 10^{-5} for the trainable encoder blocks, 5×10^{-4} for the segmentation decoder, and 10^{-4} for the height decoder. We sample training images uniformly with replacement, without augmentation, using seed 64. Every model receives 2,175 optimizer updates and is evaluated at the final checkpoint.

Loss functions. Semantic segmentation uses unweighted cross-entropy with ignore index 255. Height estimation uses a Softplus output and a weighted L1 loss over finite, non-negative targets. Predictions and targets are divided by 30 m for loss computation. Pixels with reference heights above 1 m receive weight 5, and other valid pixels receive weight 1; the weighted loss is normalized by the sum of valid pixel weights. This loss scaling does not clip metric heights.

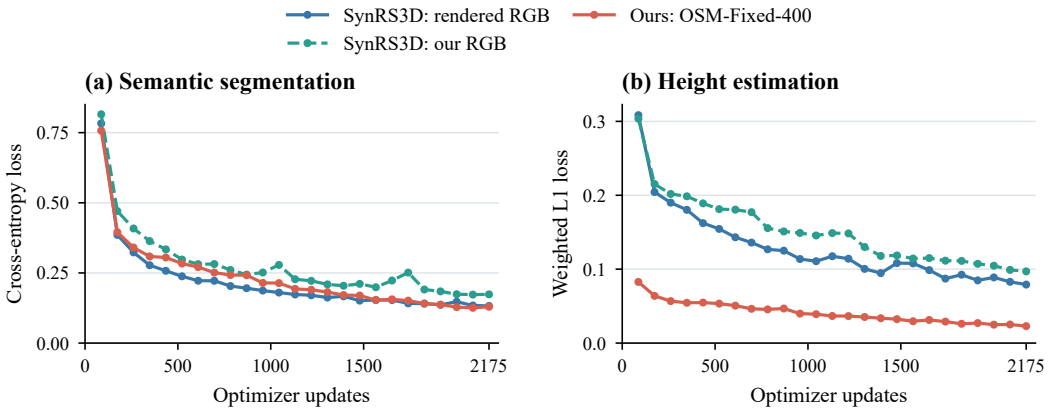


Figure 19: Downstream training losses with 400 images per source. Each point averages 87 consecutive optimizer updates; all 25 recorded intervals are shown. Segmentation uses cross-entropy, and height estimation uses the normalized weighted L1 objective described above.

Scope of the fixed-layout comparison. The fixed-layout comparison in Table 1 uses SynRS3D, which provides compatible semantic and height conditions. Exploratory RGB replacement on SMARS, SyntCities, and TopAir did not yield consistent downstream improvements. The SynRS3D

comparison therefore supports the reported result for that source and should not be interpreted as a uniform improvement across synthetic datasets.

Training dynamics. Figure 19 reports training losses for rendered SynRS3D RGB, our generated RGB on the same SynRS3D layouts, and OSM-Fixed-400. These curves describe fitting to the training source; decreasing training loss alone does not establish generalization to real imagery or other regions.

A.8 Effect of Random Seeds

Figure 20 compares six OSM layouts from Omaha and Jacksonville using four random seeds, with all other generation settings fixed. Water and vegetation show pronounced changes in color and texture. Building and road structures remain comparatively stable, with variations in roof brightness and local detail. Some grass regions change between green vegetation and a bare-soil appearance, indicating that seed variation can affect local semantic consistency as well as appearance.

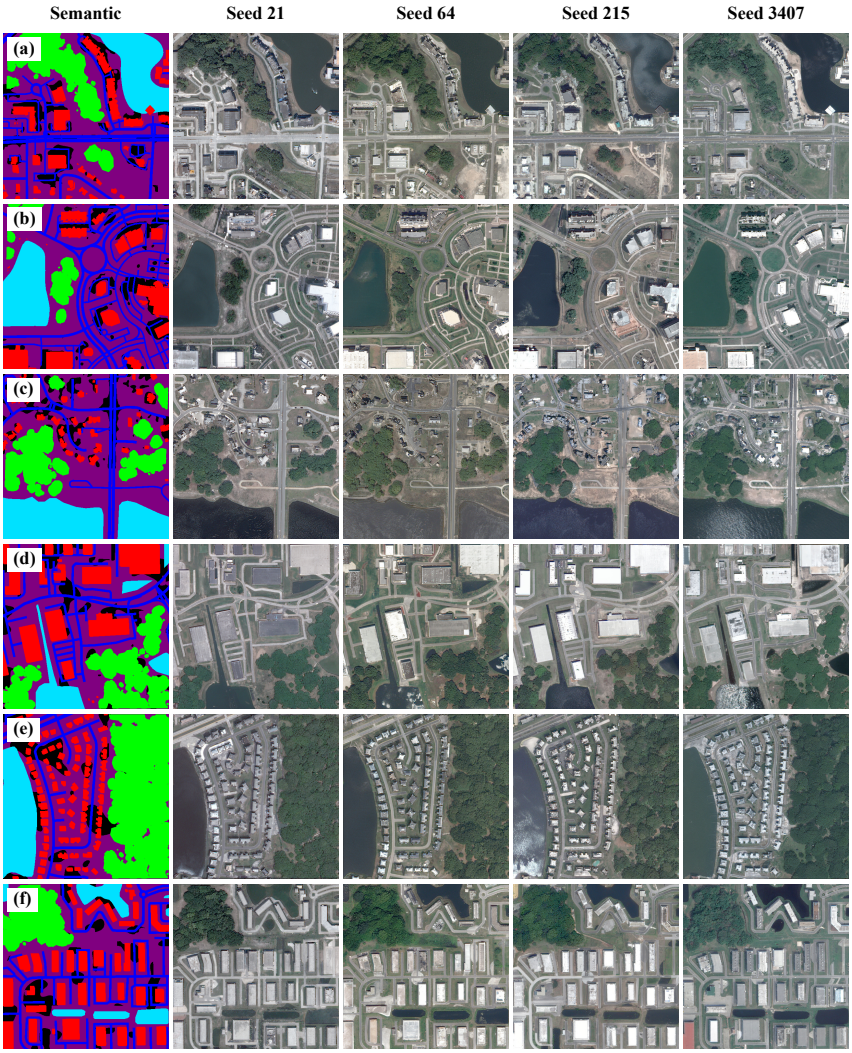


Figure 20: Random-seed variation with fixed semantic and height conditions. Rows (a)–(c) show Omaha and (d)–(f) Jacksonville. All six layouts were selected by class composition before generation. Each row uses the same e300 generator, prompt, and sampling settings; only the noise seed changes. Images are shown without color adjustment.