Diffusion Transformers for Roof Graph Synthesis and Reconstruction

Daniel Panangian and Ksenia Bittner

Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany

Abstract. We present RoofDiT, a generative framework for 2D roof graph synthesis and reconstruction. Roofs are compactly described as planar graphs of junctions and structural edges, but existing methods often rely on fixed geometric rules or direct reconstruction objectives. RoofDiT instead models roof structures directly as vertex-edge graphs and learns a conditional generative prior over their geometry and connectivity. Our framework follows a two-stage design: a diffusion transformer generates roof vertices, and an edge prediction module infers the corresponding graph topology. To improve geometric fidelity, RoofDiT combines relative geometry-aware attention with footprint and aerial-image conditioning, while using an alignment regularizer to encourage common horizontal, vertical, and diagonal roof patterns. The same model supports unconditional generation, footprint-conditioned synthesis, and image-guided reconstruction by changing the conditioning signal. Experiments show improved graph generation quality over a diffusion baseline, favorable performance against a straight-skeleton prior in the footprint-conditioned setting, and the highest edge F1 among compared methods for image-guided reconstruction.

Keywords: roof graph generation · diffusion transformer · generative modeling · aerial imagery

1 Introduction

Residential roof structures follow strong geometric and compositional regularities [17]. Their layouts are not arbitrary collections of edges and vertices, but follow recurring patterns in alignment, connectivity, and part composition. At the same time, roof structures remain diverse. Buildings with similar footprints can have different valid roof structures. This makes roof structure a natural target for generative modeling. A useful model must capture both the regularities that make roof graphs structurally valid and the diversity that allows multiple plausible configurations for the same input.

Recent years have seen substantial progress in generative modeling of structured visual and geometric data. Beyond image synthesis, generative models have been successfully applied to domains such as layouts and floorplans, where the objective is not only to produce realistic outputs but also to capture the underlying organization and constraints of the data. Roof structure reconstruction

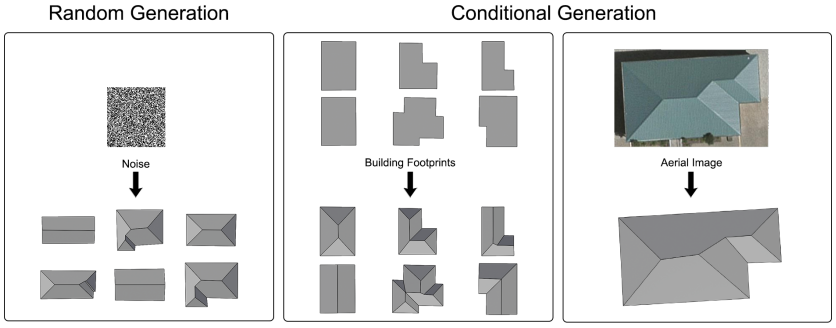


Fig. 1: Illustration of the three roof graph generation tasks in our framework: (1) unconditional random generation from noise, (2) conditional generation from building footprints, and (3) image-guided generation for reconstructing complete roof structure. The figure highlights the ability of our model to sample plausible roof layouts from a learned prior and to produce constrained predictions using geometric and visual observations.

remains challenging because available observations often provide only partial evidence about the realized geometry. Many recent methods rely primarily on image-driven cues to infer structural elements and assemble them into a roof layout [23, 21, 5]. These approaches are designed for deterministic single-output prediction and are structurally unable to represent the ambiguity inherent in the footprint-to-roof mapping. Other methods address this difficulty by incorporating stronger geometric priors or topological constraints, improving robustness when image evidence alone is insufficient, but similarly produce fixed outputs for a given input.

In this work, we investigate roof structure modeling through a generative framework that treats unconditional generation, footprint-conditioned generation, and image-guided reconstruction as three instantiations of the same conditional model under different levels of available evidence. Figure 1 provides an overview of these tasks. By studying all three settings within a single framework, we evaluate how a learned roof prior supports roof graph generation across a spectrum of conditioning from none to fully specified.

Our contributions are as follows:

- We formulate roof structure modeling as a generative problem over 2D roof graphs, motivated by the observation that the footprint-to-roof mapping is one-to-many, and demonstrate that a learned prior captures meaningful structural regularities across unconditional, footprint-conditioned, and image-guided settings.
- We design a diffusion transformer for roof graph generation that incorporates three roof-specific inductive biases: relative geometry-aware attention to capture pairwise spatial relations between roof junctions, a roof align-

ment regularizer that encourages the horizontal, vertical, and diagonal patterns common in residential roof layouts, and multimodal conditioning on building footprints and aerial imagery to support three inference settings.
– We introduce a best-of-K evaluation protocol for generative roof graph models and show it reveals a gap between representational capacity and sample selection, identifying reliable candidate selection as the key remaining challenge.

2 Related Work

2.1 Generative Models for Layouts

Generative modeling of structured spatial data has been studied extensively in floorplan, layout, and geometric design tasks. Early approaches often relied on raster or image-based formulations, generating floorplans through intermediate representations such as room masks or occupancy maps before recovering structured geometry. GAN-based methods showed that layout generation can be treated as a multimodal problem, where similar inputs may admit multiple plausible outputs [15]. More recent methods have moved toward vector and graph-based formulations that are more naturally aligned with geometric reasoning. Graph2Plan generates floorplans from layout graphs [9], while HouseGAN++ models relational structure for floorplan synthesis and refinement [15]. Diffusion-based models have further expanded this line of work. HouseDiffusion models floorplans as structured polygonal outputs given a bubble diagram specifying room count and adjacency [19], while GSDiff studies geometry-enhanced structural graph generation for vector floorplans [10]. Related work has also considered controllable layout generation [11] and vector floorplan generation under diverse conditioning signals [8]. These works show that generative models can capture both regularity and diversity in domains where outputs must satisfy nontrivial spatial and structural constraints. Our work adopts the same general perspective but applies it to roof structures, which differ from floorplans in that their internal topology is not derivable from the footprint alone and the footprint-to-roof mapping is inherently one-to-many.

2.2 Roof Structure Reconstruction from Aerial Imagery

Recovering roof structure from aerial and remote-sensing imagery has been widely studied as a structured reconstruction problem. Early work explored vectorized extraction and graph-based representations from overhead imagery using geometric reasoning and learning-based pipelines [13, 24]. More recent methods have focused directly on roof structure extraction. RSGNN reconstructs vectorized planar roof structures from very high resolution imagery by combining image evidence with graph reasoning [23]. HEAT framed structured reconstruction as transformer-based planar graph prediction from raster inputs [5]. Segmentation-based approaches have also been explored alongside these structured prediction

methods [22]. PolyRoof extended polygonal graph-based extraction to residential roof structure prediction [2], and RoofMapNet formulates roof recovery as structured prediction over geometric primitives from high-resolution imagery [21]. These methods are closest to our work in application domain, but they are designed for deterministic reconstruction from image evidence, producing a single prediction per input and offering no mechanism for representing structural ambiguity or sampling alternative plausible configurations. Our work studies roof prediction through a generative framework that naturally supports multiple outputs under the same conditioning signal.

2.3 Roof Modeling and Generation

A smaller but important body of work considers roofs directly as objects of modeling or generation rather than only reconstruction targets. Much of this literature comes from computer graphics, where the focus is on 3D roof construction from building outlines or procedural rules. Classical approaches based on straight skeletons generate roof structures from footprints under geometric assumptions [1, 3, 6, 12]. Related procedural and grammar-based methods model roofs through combinations of elementary primitives or rule-based constructions [14, 4]. More recent work introduced roof-specific graph representations and optimization-based refinement for planar roof modeling [18]. Learning-based roof generation has received comparatively less attention. Roof-GAN models residential roofs as compositions of roof primitives and their relations [17], highlighting the value of learning roof-specific structural priors rather than treating roof synthesis as unstructured generation. Related graph-constrained generative approaches have also considered roof generation as part of broader house generation settings [20]. In contrast to these works, which are primarily concerned with 3D roof construction or primitive-based synthesis, our work focuses on planar vertex-edge roof graphs as a 2D generative structural representation.

3 Dataset

We use the residential roof dataset introduced by [18] as the basis for our experiments. The dataset contains annotated residential roofs together with corresponding aerial images, and is split into 1,926 training samples, 249 validation samples, and 223 test samples. We use this dataset because it provides a moderate-scale collection of residential roof structures with both geometric annotations and image observations, making it suitable for studying graph generation and image-guided prediction.

From the roof annotations, we derive 2D roof graphs in top view, where vertices correspond to roof junctions and edges correspond to roofline segments. Before training, we canonicalize the annotations to reduce unnecessary variation in orientation. Specifically, each roof is rotated so that the longest footprint edge aligns with a fixed reference axis. We then regularize edge directions by snapping nearly horizontal and vertical segments to the corresponding axis-aligned

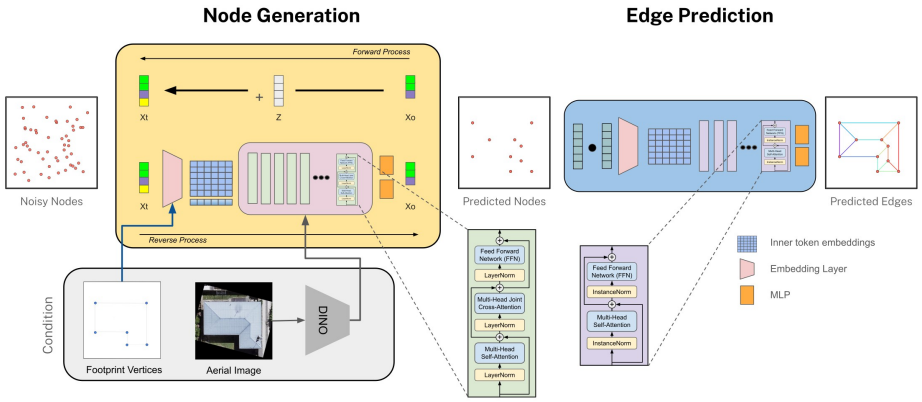


Fig. 2: Overall pipeline used in our method. Following GSDiff, the framework decomposes roof graph generation into two stages: node generation and edge prediction. In the first stage, a conditional diffusion transformer starts from noisy node states and iteratively denoises them to generate roof vertices, optionally conditioned on footprint vertices, pretrained aerial image features, or both. In the second stage, an edge prediction module takes the generated node set and infers pairwise connectivity to produce the final roof graph.

directions, which reduces small annotation inconsistencies while preserving the overall roof structure. For the footprint-conditioned setting, we additionally derive footprint vertices from the roof graph by taking the exterior boundary of the union of planar roof faces, and use these vertices as conditioning input.

4 Approach

4.1 Problem Formulation

We represent a roof as a planar graph $G = (V, E)$ in top view, where V is a set of roof junction vertices with 2D coordinates and E is a set of roofline edges connecting them. A roof graph is considered valid if it is non-empty, planar, free of dangling degree-one nodes, and contains at least one cycle. We model roof graph generation as learning a conditional distribution

$$p_\theta(G \mid c), \tag{1}$$

where c is a conditioning signal that may be empty (unconditional), a set of building footprint vertices (footprint-conditioned), aerial image features (image-guided), or both. By varying c, the same model supports three inference settings without architectural changes.

4.2 Overview

We build on the two-stage graph generation framework of GSDiff [10], which generates nodes first and then predicts edges. We keep this pipeline across all three settings and modify the node generation stage with relative geometry-aware attention, a roof alignment regularizer and multimodal conditioning.

4.3 Node Generation

Let $X \in \mathbb{R}^{N\times d}$ denote the node state of a roof graph, where N is the maximum number of nodes and d includes 2D coordinates and a validity flag. Node generation follows the DDPM framework [7] as implemented in GSDiff [10]: the model is trained to predict noise ϵ from noisy states X_t, and at inference time denoises from Gaussian noise to roof vertex positions.

Relative geometry-aware attention Self-attention over node tokens does not explicitly encode the geometric relations that are central to roof structure. To incorporate this information, we augment attention with pairwise relative geometry features. For two nodes i and j, we compute

$$\phi_{ij} = [\Delta x_{ij}, \Delta y_{ij}, r_{ij}, u_{x,ij}, u_{y,ij}], \tag{2}$$

where Δx_{ij} and Δy_{ij} are coordinate offsets, r_{ij} is the Euclidean distance, and $(u_{x,ij}, u_{y,ij}) = (\Delta x_{ij}, \Delta y_{ij})/r_{ij}$ is the unit direction from node i to node j. These features are mapped by a small MLP to produce per-head attention biases:

$$A_{ij} \leftarrow A_{ij} + \mathrm{MLP}(\phi_{ij}), \tag{3}$$

where the MLP outputs one scalar per attention head per node pair, so token interactions depend on both learned content and explicit geometric relations between roof vertices.

Roof alignment Roof layouts often contain horizontal, vertical, and diagonal structural alignments. We define a lightweight alignment regularizer over four canonical undirected line families:

$$\begin{cases} x = \text{const}, & y = \text{const}, \\ x + y = \text{const}, & x - y = \text{const}. \end{cases} \tag{4}$$

The per-pair distances to each line family are:

$$d^x_{ij} = |x_i - x_j|, \quad d^y_{ij} = |y_i - y_j|, \quad d^+_{ij} = \tfrac{|(x_i+y_i)-(x_j+y_j)|}{\sqrt{2}}, \quad d^-_{ij} = \tfrac{|(x_i-y_i)-(x_j-y_j)|}{\sqrt{2}}. \tag{5}$$

For each valid node i, we measure its smallest support-line distance to any other valid node:

$$m_i = \min_{j\neq i} \min\left(d^x_{ij}, d^y_{ij}, d^+_{ij}, d^-_{ij}\right). \tag{6}$$

These distances are converted into penalties with time-dependent weighting over diffusion steps. The final node generation objective combines the diffusion loss with the roof alignment term:

$$\mathcal{L}_{\text{node}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{align}} \mathcal{L}_{\text{roof-align}} \tag{7}$$

Conditioning design We use two complementary conditioning modalities. For footprint-conditioned generation, footprint vertices are encoded as 2D geometric tokens in the same representation space as the generated roof nodes and concatenated to the node sequence, so the transformer processes roof nodes and footprint vertices jointly through self-attention. This enables reasoning over interactions between the known outer building geometry and the internal roof structure. For image-guided reconstruction, features extracted by a pretrained DINOv2 [16] encoder are linearly projected into a sequence of image feature tokens $M \in \mathbb{R}^{P \times d}$ and injected through cross-attention at each transformer block. The two conditioning sources can also be combined.

4.4 Edge Prediction

After node generation, the second stage predicts graph connectivity from the generated node set. Given the generated roof vertices, the edge prediction module evaluates candidate node pairs and predicts whether an edge should connect them, yielding the final roof graph G. We keep the edge prediction stage unchanged from the underlying two-stage graph pipeline.

5 Experiments

5.1 Experimental Setup

We use the train/validation/test split described in Sec. 3. We evaluate three settings: unconditional generation, footprint-conditioned generation, and image-guided reconstruction. In all cases, RoofDiT uses two-stage prediction with node generation followed by edge prediction. For footprint-conditioned generation, the model is conditioned on the building footprint. For image-guided reconstruction, the model is conditioned on aerial image features alone or in combination with the building footprint. In the generative settings, inference is stochastic, and for footprint-conditioned generation we sample five outputs per test footprint and report averaged and best-of-5 metrics where appropriate. For image-guided reconstruction, inference is deterministic and uses a single prediction per input. All models are trained with AdamW for 100,000 steps with batch size 128, learning rate 10^{-4}, and a cosine diffusion schedule of 1000 steps. The alignment loss weight is set to $\lambda_{\text{align}} = 1$. For image-guided reconstruction, aerial images are encoded using a frozen pretrained DINOv2 ViT-S/14 backbone. The node generation transformer has 24 layers, hidden dimension d=512, and 4 attention heads; the geometry bias MLP has hidden dimension 64. The maximum node count is N=53 and coordinates are normalized to $[0, 1]$.

5.2 Compared Methods

We compare RoofDiT with GSDiff for unconditional generation and straight skeleton for footprint-conditioned generation. For image-guided reconstruction, we use HEAT and RoofMapNet as the main comparison methods. HEAT is retrained on the same split as RoofDiT, while RoofMapNet is evaluated using its released pretrained model. HEAT and RoofMapNet use pixel-aligned roof annotations whereas RoofDiT uses canonicalized graph representations.

5.3 Evaluation Metrics

We use different protocols for the three settings. For unconditional generation, we generate 1000 roof graphs and compare them against all 223 unique test samples using FID and KID computed from InceptionV3 (2048-dim) features. We also report Roof Valid Rate and Planar Rate.

For footprint-conditioned generation, five samples are generated per test footprint. Since a footprint may correspond to multiple plausible roofs, we report best-of-5 node, edge, and face metrics. Geometry metrics such as Valid Rate and Planar Rate are averaged over all samples. For image-guided reconstruction, all metrics are computed on a single prediction.

A roof graph $G = (V, E)$ is considered valid if it is non-empty, planar, free of dangling degree-one nodes, and contains at least one cycle. Roof Valid Rate and Planar Rate are defined as

$$\text{RoofValidRate} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{G_i \text{ is valid}\},$$

$$\text{PlanarRate} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{G_i \text{ has no edge crossings}\}.$$

Node evaluation uses Hungarian matching under a spatial threshold, from which we compute precision, recall, F1, and node count MAE. Edge evaluation counts an edge as correct when both endpoints are matched and connectivity is preserved. Face evaluation follows the HEAT region-based protocol [5], reporting precision, recall, F1, matched IoU, and face count error. We additionally report the number of components, crossings, and dangling nodes as geometry metrics, where lower values indicate better graph quality. Unless otherwise stated, node and edge metrics are reported at a 5-pixel threshold.

6 Results

6.1 Unconditional Generation

Table 1 compares RoofDiT against the GSDiff baseline on unconditional roof graph generation. RoofDiT improves all metrics, with clearly lower FID and KID indicating better distributional realism and diversity, and slightly higher valid

Table 1: Unconditional roof graph generation.

Model	FID ↓	KID ↓	Valid ↑	Planar ↑
GSDiff	49.4	57.4	0.975	0.977
RoofDiT	**34.6**	**31.7**	**0.985**	**0.988**

Table 2: Performance on footprint-conditioned roof graph generation. SS denotes the straight skeleton baseline.

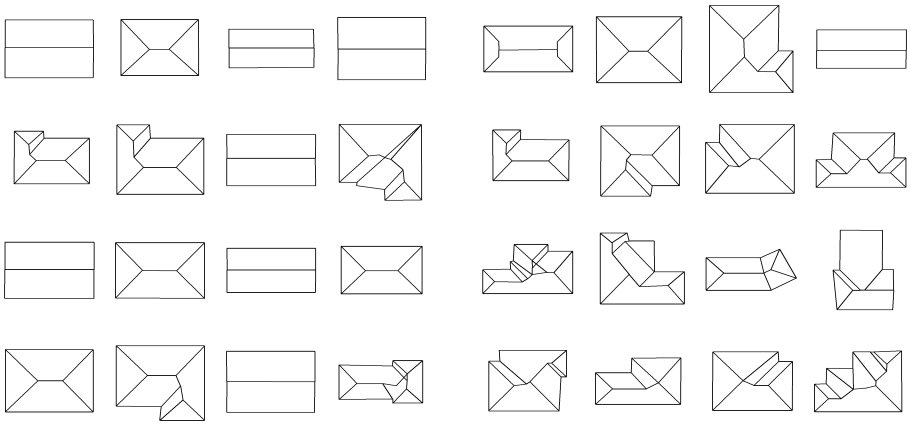
	Metric	SS	RoofDiT
Node	Count MAE ↓	2.008	**1.002**
	F1@5 ↑	**0.849**	0.705
Edge	F1 ↑	0.960	**0.981**
Face	Count Err. ↓	1.430	**0.444**
	F1 ↑	**0.840**	0.798
	$F1_{\text{best-of-}K}$ ↑	0.840	**0.886**
	Matched IoU ↑	**0.850**	0.697
Geometry	Planar Rate ↑	**1.000**	0.873
	Valid Rate ↑	**1.000**	0.841
	$\text{Valid Rate}_{\text{best-of-}K}$ ↑	**1.000**	0.930

and planar rates. Both methods reliably generate structurally sound graphs; the larger gains in FID and KID reflect improvements in the range and quality of generated layouts. These trends are visible in Fig. 3: GSDiff tends to produce simpler, more repetitive layouts dominated by symmetric ridge patterns, whereas RoofDiT generates a broader range of configurations, including more complex asymmetric roofs.

6.2 Footprint-Conditioned Generation

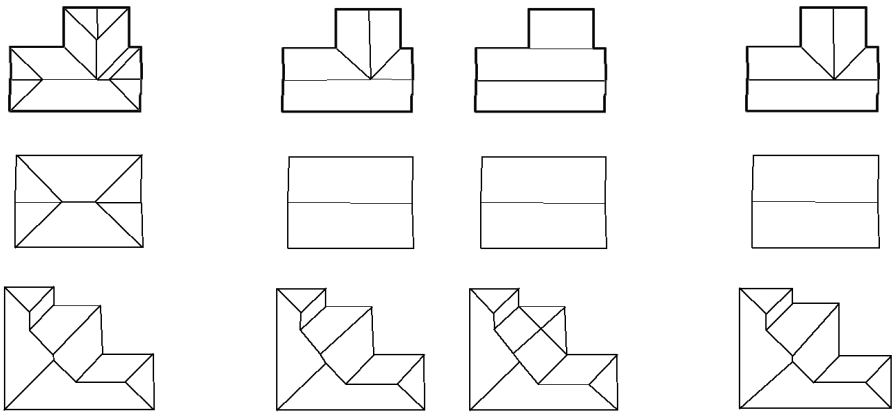
Table 2 compares footprint-conditioned generation between the straight skeleton baseline and RoofDiT. RoofDiT achieves substantially lower node count MAE and face count error, and improves edge F1 from 0.960 to 0.981. Under best-of-K evaluation, it further improves face F1 from 0.798 to 0.886 and valid rate from 0.841 to 0.930. These results indicate that RoofDiT better recovers the overall amount of roof structure and benefits from sampling multiple plausible candidates for the same footprint. At the same time, the straight skeleton baseline remains stronger on node F1@5, face F1, matched IoU, planar rate, and valid rate, reflecting its handcrafted geometric prior and deterministic structural regularity.

The qualitative examples in Fig. 4 illustrate this tradeoff clearly. Straight skeleton can only return a fixed procedural interpretation of the footprint, often overpredicting nodes and faces through unnecessary internal partitions.



(a) GSDiff (b) RoofDiT

Fig. 3: Qualitative comparison of unconditional roof graph generation. Samples produced by GSDiff and RoofDiT are shown side by side.



(a) SS (b) RoofDiT (c) GT

Fig. 4: Qualitative comparison for footprint-conditioned roof graph generation across three building examples. For each row, the straight skeleton (SS) produces a single deterministic output from the footprint. RoofDiT is able to generate multiple plausible outputs per footprint, illustrating the diversity of plausible roof configurations consistent with the same outer boundary.

RoofDiT, in contrast, often produces samples that are closer to the ground truth in ridge layout and face decomposition, while also generating meaningfully different roof organizations for the same footprint rather than minor perturbations

of a single solution. This diversity is meaningful precisely because a building footprint does not uniquely determine a single roof topology: the best-of-K gains show that good solutions exist within the model's sample distribution, and the primary remaining challenge is selecting them reliably from a single draw.

6.3 Image-Guided Reconstruction

Table 3 compares image-guided roof graph reconstruction across node, edge, face, and geometry metrics. It is important to note that HEAT and RoofMapNet are discriminative reconstruction models trained end-to-end to minimize annotation error on a single prediction, while RoofDiT is a generative model evaluated on reconstruction as one of three supported inference modes. This difference in design objective is reflected in the results.

RoofMapNet performs substantially worse than the other methods across all metric groups. HEAT achieves the strongest overall performance, with the best node precision, recall, and F1, the strongest face metrics, and the best geometry scores. For RoofDiT, adding footprint conditioning improves performance substantially over using aerial imagery alone: node F1 increases from 42.6 to 70.3, edge F1 from 78.4 to 98.4, face F1 from 63.5 to 82.4, and node count MAE decreases from 2.23 to 0.90. When both aerial imagery and building footprints are provided, RoofDiT achieves the strongest edge precision, recall, and F1 among all compared methods, which is notable because edge structure carries much of the topological information in a roof graph.

The qualitative examples in Fig. 5 complement the quantitative results. In the first row, RoofDiT reconstructs a plausible roof graph with a clean overall structure, though HEAT is slightly closer to the ground truth in some local details. In the second row, RoofDiT performs best among the compared methods, capturing the main topology more faithfully while the other methods fail to recover the simple target structure. In the third row, RoofDiT recovers a plausible roof family and the main global layout on a more articulated example. In the fourth row, RoofDiT simultaneously over-generates nodes in the central region, causing slanted and irregular ridge lines, and under-segments the right section, producing fewer faces than the ground truth. The predicted graph also contains overlapping edges. Overall, RoofDiT can produce strong and geometrically regular predictions, but does not always match the ground truth as closely as the reconstruction baseline.

6.4 Ablation Study

We ablate the two main components added over GSDiff in the unconditional setting, where FID and KID provide a direct distributional signal of generation quality. Table 4 shows that relative geometry-aware attention is the primary driver of improvement, reducing FID from 49.4 to 37.1 and KID from 57.4 to 36.5. The roof alignment regularizer provides a further consistent gain across all

Table 3: Comparison of image-conditioned roof graph reconstruction methods grouped by node, edge, face, and geometry metrics. Node metrics are reported at a 5-pixel matching threshold. $\text{RoofDiT}_{\text{aerial}}$ uses aerial imagery only, while $\text{RoofDiT}_{\text{aerial+footprint}}$ additionally uses the building footprint as input.

	Node				Edge			Face					Geometry				
Method	Prec	Recall	F1	Count MAE↓	Prec	Recall	F1	Prec	Recall	F1	Matched IoU	Count Err.↓	Valid	Planar	#Comp↓	Crossings↓	Dangling↓
RoofMapNet	27.9	40.8	32.7	7.40	43.6	45.9	44.5	25.5	27.8	26.1	25.9	0.30	20.2	38.1	5.12	3.22	1.51
HEAT	**93.3**	**93.0**	**93.0**	**0.30**	93.9	94.2	94.0	**91.6**	**89.9**	**90.2**	**86.2**	**0.11**	**90.6**	**92.8**	**1.00**	**0.11**	0.09
$\text{RoofDiT}_{\text{aerial}}$	42.7	43.9	42.6	2.23	78.1	80.5	78.4	63.7	66.2	63.5	60.4	1.12	87.4	90.1	**1.00**	0.13	**0.04**
$\text{RoofDiT}_{\text{aerial+footprint}}$	70.5	70.9	70.3	0.90	**98.5**	**98.5**	**98.4**	81.9	83.8	82.4	71.8	0.48	88.3	90.6	**1.00**	0.19	**0.04**

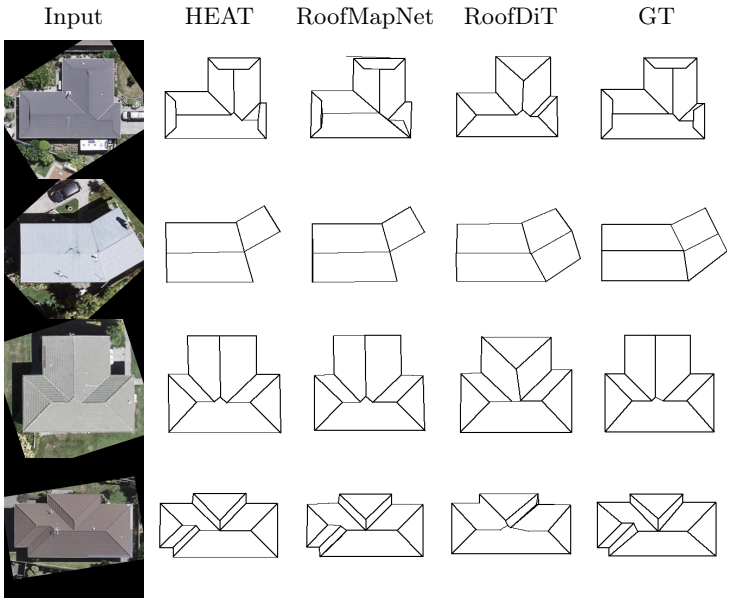


Fig. 5: Qualitative comparison of image-guided roof graph reconstruction. RoofDiT is conditioned on the input aerial image only, without building footprint. HEAT and RoofMapNet are shown for comparison against the ground-truth roof graph.

Table 4: Ablation study on unconditional generation.

Model	FID ↓	KID ↓	Valid ↑	Planar ↑
GSDiff (baseline)	49.4	57.4	0.975	0.977
+Rel. geometry-aware attention	37.1	36.5	0.977	0.983
+Roof alignment	**34.6**	**31.7**	**0.985**	**0.988**

metrics, with the full RoofDiT model reaching FID 34.6 and KID 31.7. Both components contribute independently.

7 Discussion

The results suggest that roof structure can be modeled effectively as a learned prior over planar roof graphs. Across all three settings, RoofDiT captures structural regularities that remain useful under conditioning. This is especially visible in the unconditional setting, where the model produces valid and diverse roof graphs that better match the target distribution than the baseline, indicating that it learns meaningful regularities of roof topology rather than relying only on conditioning signals at inference time.

The footprint-conditioned setting reveals a particularly informative trade-off. Geometric constructions such as straight skeleton offer strong guarantees of validity and regularity but are constrained to a single deterministic output per footprint. A learned generative model, by contrast, can represent a broader range of plausible roof organizations consistent with the same footprint. The best-of-K gains are the clearest evidence of this: face F1 improves from 0.798 to 0.886 and valid rate from 0.841 to 0.930 when selecting the best of five samples. This shows that good solutions exist within the model's sample distribution, and the primary remaining challenge is sample selection, not representational capacity.

The image-guided results follow a similar pattern. HEAT, as a discriminative model, achieves higher precision on node and face metrics. RoofDiT is particularly strong on edge recovery, which carries much of the roof's topological structure, and benefits substantially from footprint conditioning when available. This suggests that the generative formulation is most valuable when visual evidence alone is insufficient and structural ambiguity is high.

Several limitations remain. Validity is encouraged through learning rather than guaranteed by construction, which leaves geometric baselines stronger in strict validity and deterministic consistency. Incorporating explicit planarity constraints during decoding is a promising direction. On the other hand, our experiments are conducted on a geographically localized dataset. Evaluating the framework across broader architectural styles, roof typologies, and regional construction patterns is an important direction for future work.

8 Conclusion

This work shows that generative roof graph modeling is a promising approach for both roof graph generation and evidence-guided roof reconstruction. Across unconditional and footprint-conditioned generation, RoofDiT demonstrates that a diffusion-based model can learn a meaningful prior over plausible planar roof graphs. In image-guided reconstruction, the method remains competitive with strong baselines and effectively exploits footprint information when available. The best-of-K results across settings consistently show that the model places good solutions within its sample distribution, pointing to sample selection as the key direction for future improvement.

Overall, the results suggest that learned generative modeling provides a useful complement to deterministic geometric methods. While geometric baselines offer

stronger guarantees in validity and consistency, the generative approach is better suited to representing multiple plausible roof graphs when the available evidence does not uniquely determine the solution. This makes it a promising foundation for future work on more reliable, structurally constrained, and geometrically richer roof reconstruction models.

References

1. Aichholzer, O., Aurenhammer, F.: Straight skeletons for general polygonal figures in the plane. In: Proceedings of the International Conference on Computing and Combinatorics. pp. 117–126 (1996)
2. Amrullah, C., Panangian, D., Bittner, K.: Polyroof: Precision roof polygonization in urban residential building with graph neural networks. In: Joint Urban Remote Sensing Event. pp. 1–4 (2025)
3. Biedl, T., Held, M., Huber, S., Kaaser, D., Palfrader, P.: Weighted straight skeletons in the plane. Computational Geometry **48**(2), 120–133 (2014)
4. Buron, C., Marvie, J.E., Gautron, P.: GPU Roof Grammars. In: Eurographics Short Papers (2013)
5. Chen, J., Qian, Y., Furukawa, Y.: Heat: Holistic edge attention transformer for structured reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3856–3865 (2022)
6. Eppstein, D., Erickson, J.: Raising roofs, crashing cycles, and playing pool: applications of a data structure for finding pairwise interactions. In: Proceedings of the Fourteenth Annual Symposium on Computational Geometry. pp. 58–67 (1998)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
8. Hong, S., Zhang, X., Du, T., Cheng, S., Wang, X., Yin, J.: Cons2plan: Vector floorplan generation from various conditions via a learning framework based on conditional diffusion models. In: Proceedings of the ACM International Conference on Multimedia. pp. 3248–3256 (2024)
9. Hu, R., Huang, Z., Tang, Y., Van Kaick, O., Zhang, H., Huang, H.: Graph2plan: Learning floorplan generation from layout graphs. ACM Transactions on Graphics **39**(4), 118:1–118:14 (2020)
10. Hu, S., Wu, W., Wang, Y., Xu, B., Zheng, L.: Gsdiff: synthesizing vector floorplans via geometry-enhanced structural graph generation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
11. Inoue, N., Kikuchi, K., Simo-Serra, E., Otani, M., Yamaguchi, K.: Layoutdm: Discrete diffusion model for controllable layout generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10167–10176 (2023)
12. Kelly, T., Wonka, P.: Interactive architectural modeling with procedural extrusions. ACM Transactions on Graphics **30**(2), 14:1–14:15 (2011)
13. Li, Z., Wegner, J.D., Lucchi, A.: Topological map extraction from overhead images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1715–1724 (2019)
14. Müller, P., Wonka, P., Haegler, S., Ulmer, A., Van Gool, L.: Procedural modeling of buildings. In: Proceedings of ACM SIGGRAPH. pp. 614–623 (2006)

15. Nauata, N., Hosseini, S., Chang, K.H., Chu, H., Cheng, C.Y., Furukawa, Y.: House-gan++: Generative adversarial layout refinement network towards intelligent computational agent for professional architects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13627–13636 (2021)
16. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
17. Qian, Y., Zhang, H., Furukawa, Y.: Roof-gan: Learning to generate roof geometry and relations for residential houses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2796–2805 (2021)
18. Ren, J., Zhang, B., Wu, B., Huang, J., Fan, L., Ovsjanikov, M., Wonka, P.: Intuitive and efficient roof modeling for reconstruction and synthesis. ACM Transactions on Graphics **40**(6), 249:1–249:17 (2021)
19. Shabani, M.A., Hosseini, S., Furukawa, Y.: Housediffusion: Vector floorplan generation via a diffusion model with discrete and continuous denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5466–5475 (2023)
20. Tang, H., Shao, L., Sebe, N., Van Gool, L.: Graph transformer gans with graph masked modeling for architectural layout generation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(6), 4298–4313 (2024)
21. Wang, J., Chen, G., Zhang, X., Wang, T., Tan, X., Yang, Q., Zhou, W., Zhu, K.: Roofmapnet: Utilizing geometric primitives for depicting planar building roof structure from high-resolution remote sensing imagery. International Journal of Applied Earth Observation and Geoinformation **141**, 104630 (2025)
22. Xu, Y., Schuegraf, P., Bittner, K.: Multi-branch convolutional neural network in building polygonization using remote sensing images. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science **93**(1), 79–100 (2025)
23. Zhao, W., Persello, C., Stein, A.: Extracting planar roof structures from very high resolution images using graph neural networks. ISPRS Journal of Photogrammetry and Remote Sensing **187**, 34–45 (2022)
24. Zorzi, S., Bazrafkan, S., Habenschuss, S., Fraundorfer, F.: Polyworld: Polygonal building extraction with graph neural networks in satellite images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1848–1857 (2022)