



Pyramid Diffusion for Fine 3D Large Scene Generation

Yuheng Liu^{1,2}

Xinke Li³

Xueting Li⁴

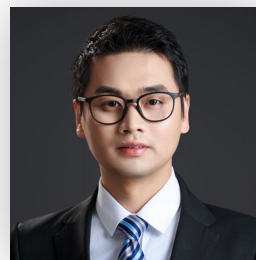
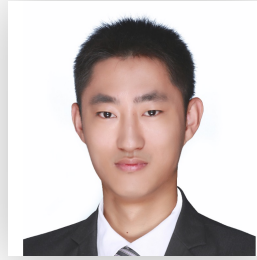
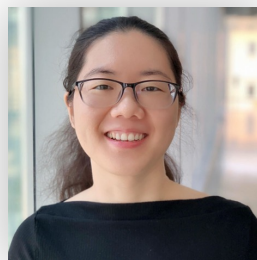
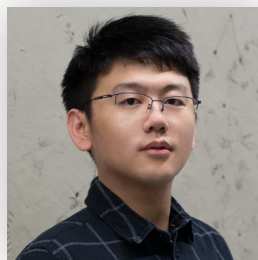
Lu Qi⁵

Chongshou Li¹

Ming-Hsuan Yang^{5,6}

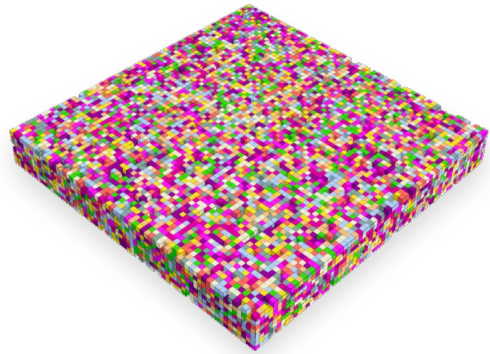
¹ Southwest Jiaotong University ² University of Leeds ³ City University of Hong Kong

⁴ NVIDIA ⁵ University of California, Merced ⁶ Yonsei University

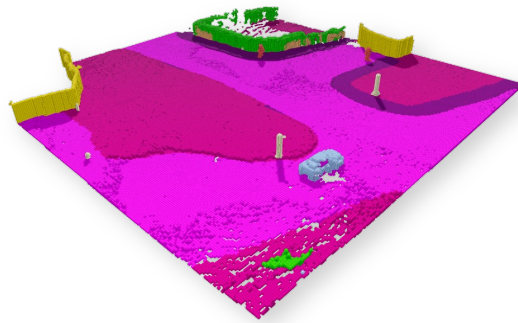


3D Large Scene Generation










Aiming to generate detailed 3D scenes.

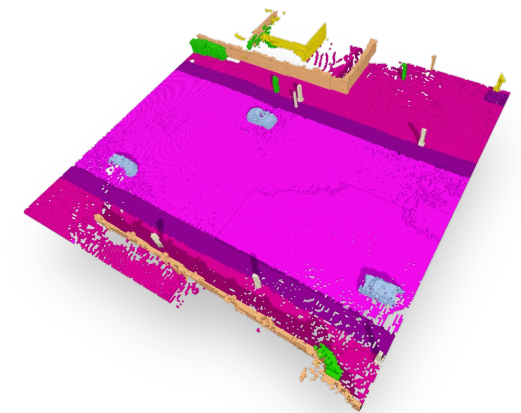
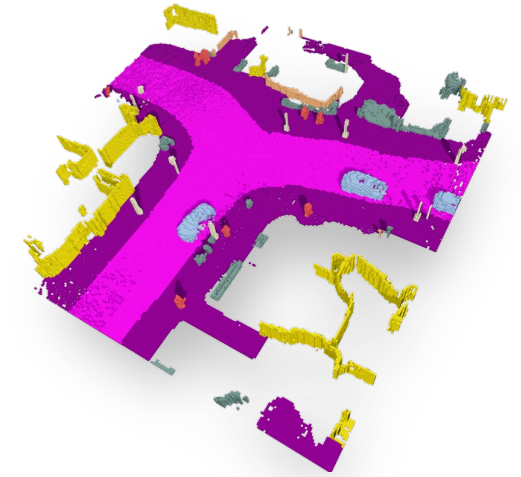
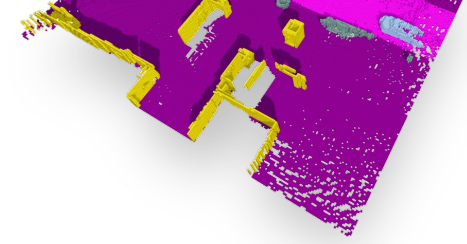


Noise



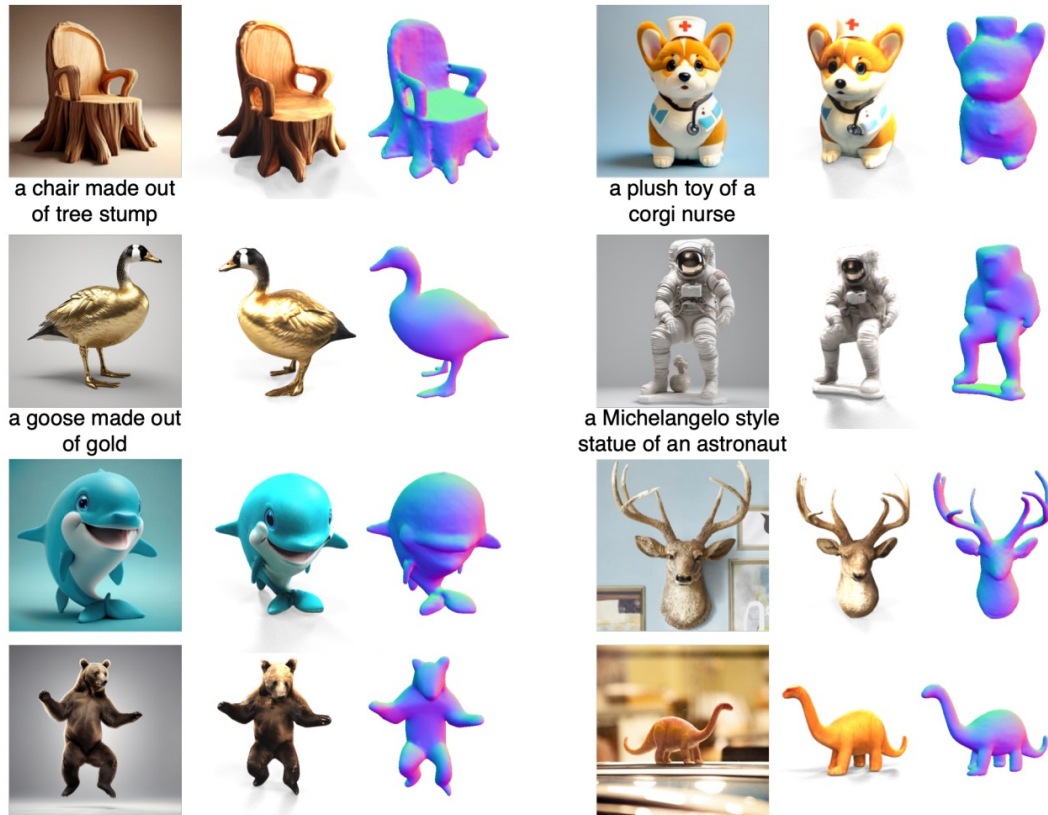
3D Large Scene
(Voxel-based)

-  Vehicles
-  Vegetation
-  Sidewalk
-  Ground
-  Pedestrians
-  Road
-  Pole
-  Barrier
-  Building
-  Other



Related Work and Limitations

Existing methods focus on single objects or indoor scene generation.



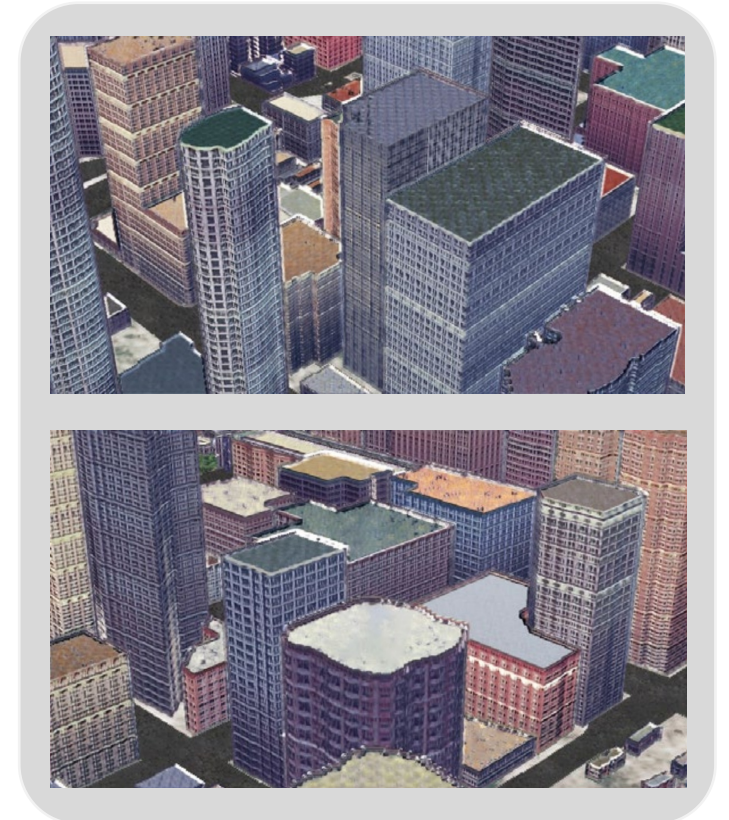
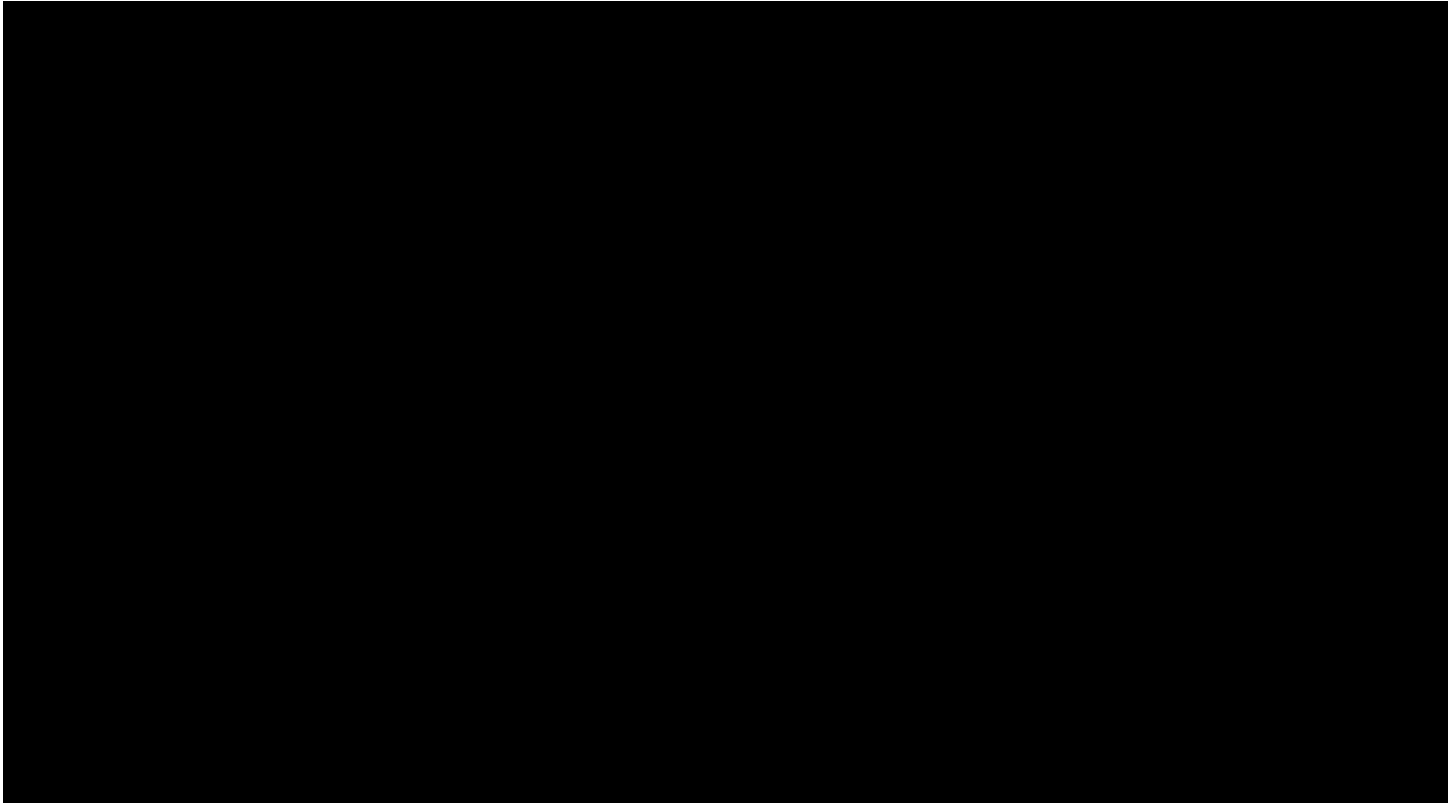
(One-2-3-45: Shi *et al* 2024)



(EchoScene: Zhai *et al* 2024)

Related Work and Limitations

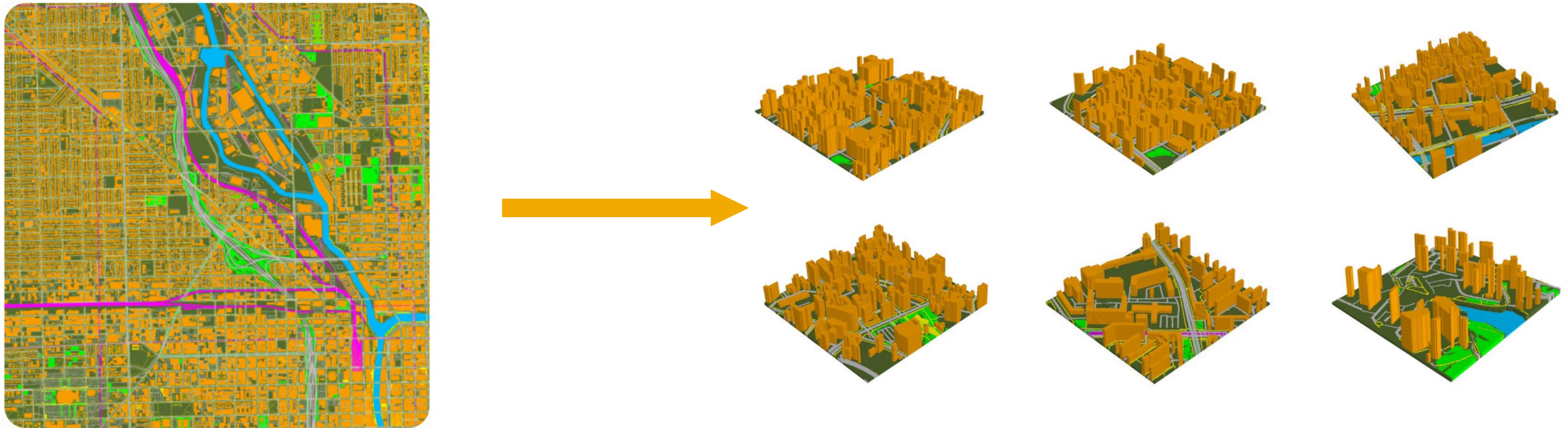
Existing methods cannot generate large 3D scenes with intricate details.



(CityDreamer: Xie *et al* 2024)

Related Work and Limitations

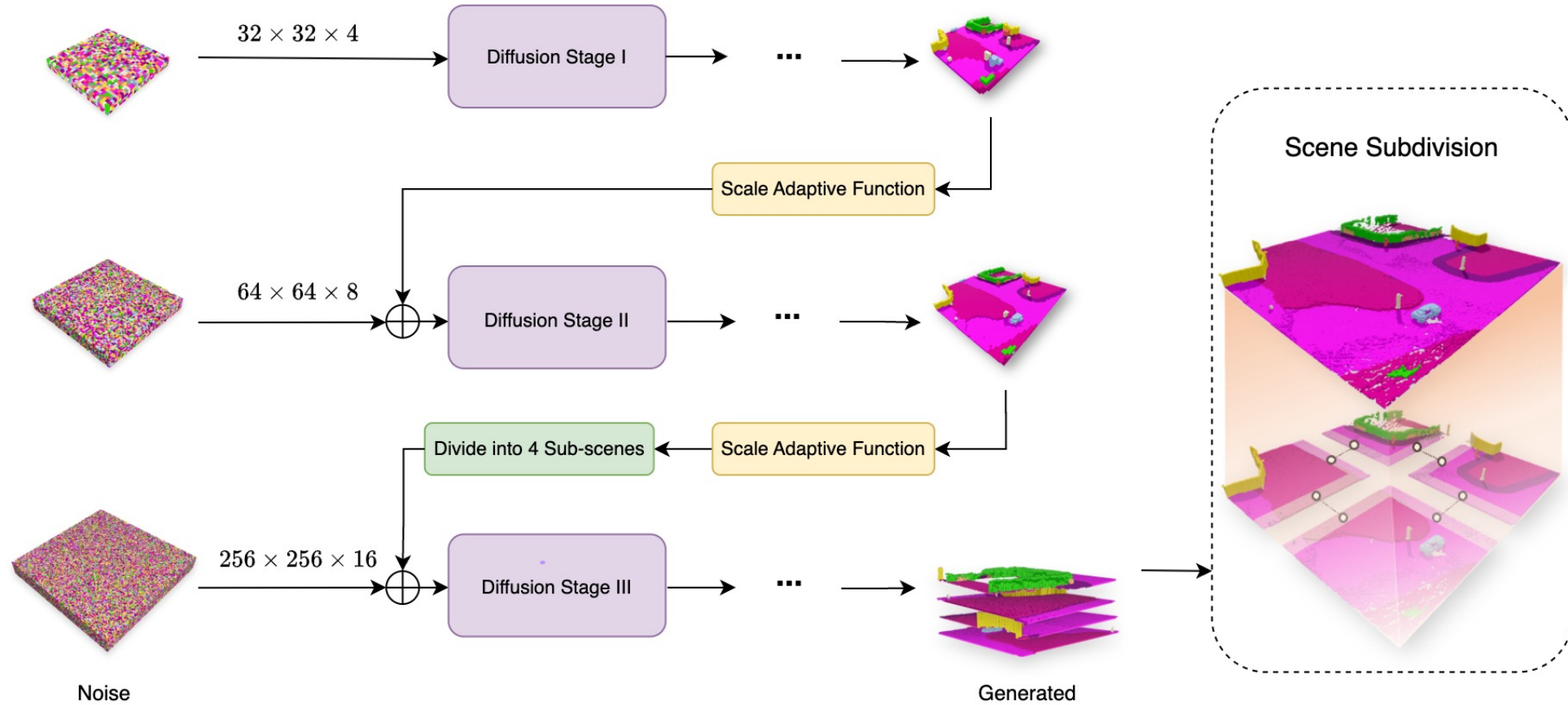
Using conditional data may limit the model's generalization ability.



(CityGen: Deng *et al* 2023)

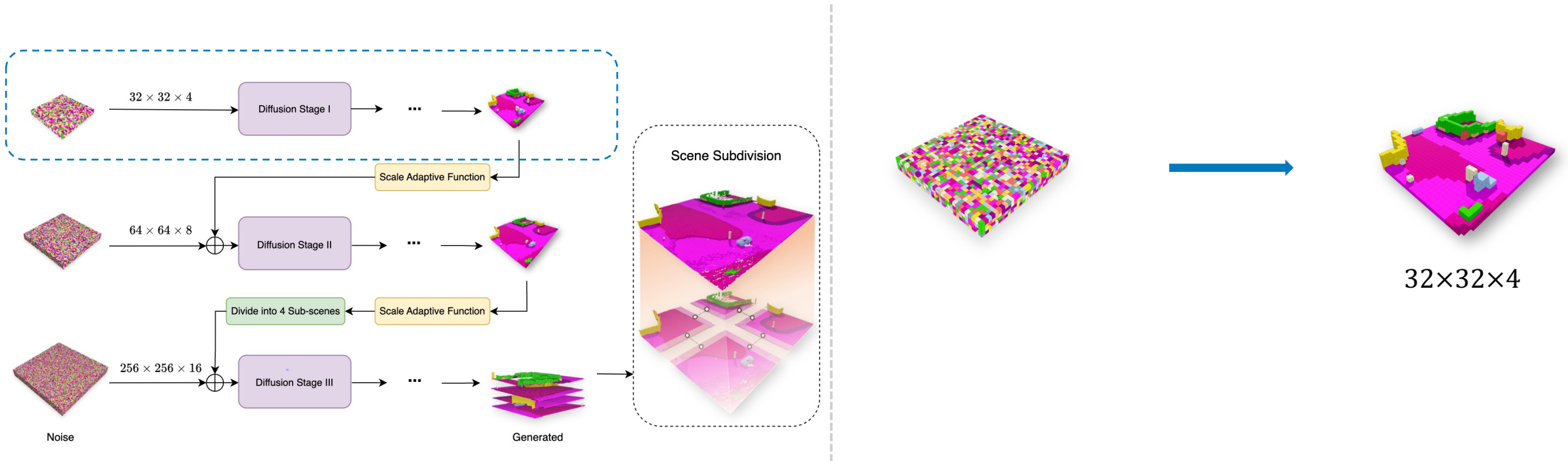
Method

Pyramid Diffusion with Scale Adaptive Function and Scene Subdivision



Method

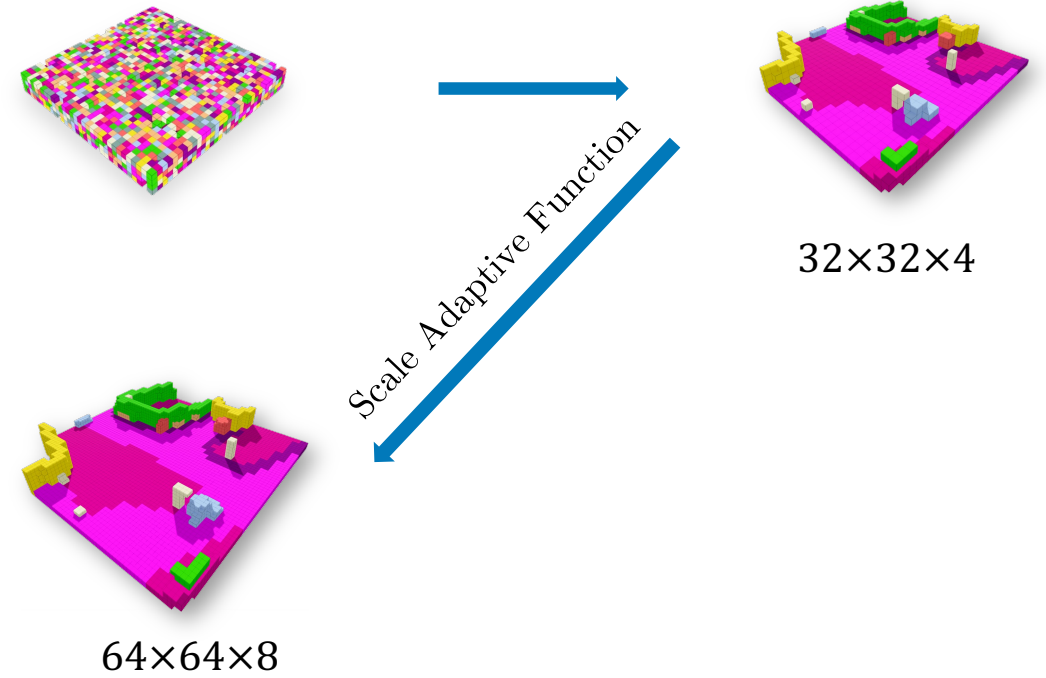
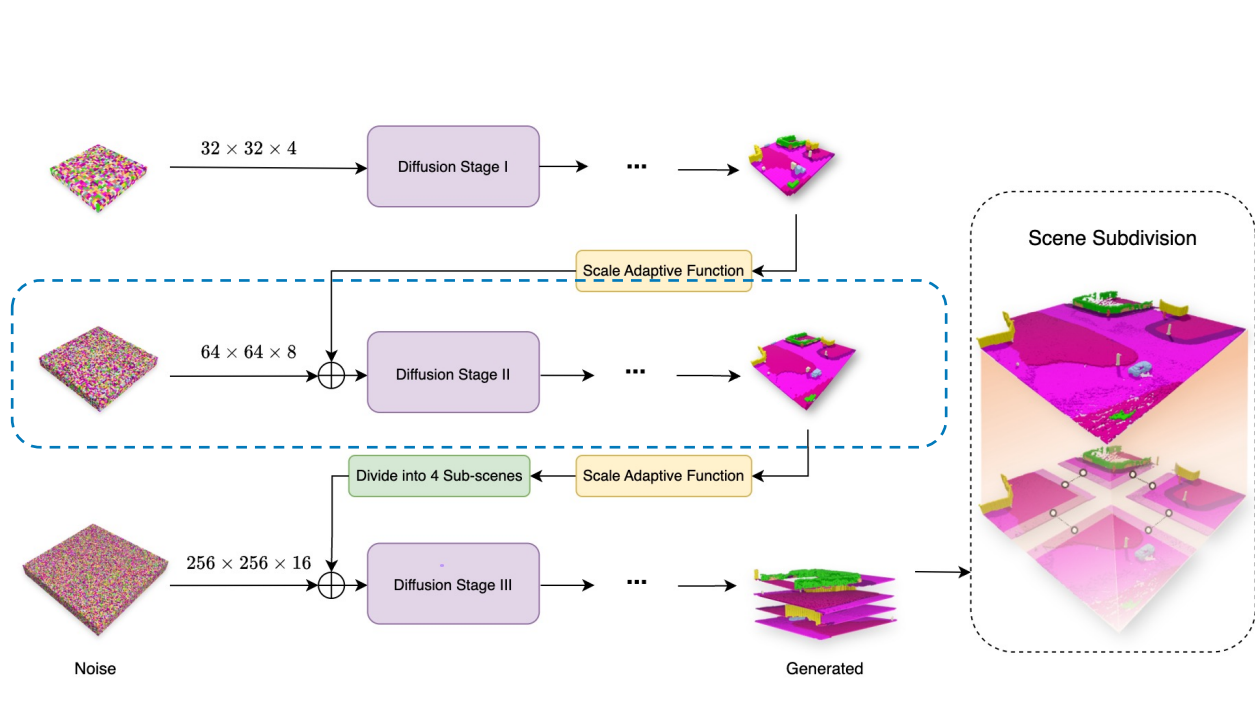
Stage I – Generation



Discrete Diffusion Generation: Random Noise \rightarrow Coarse Scene

Method

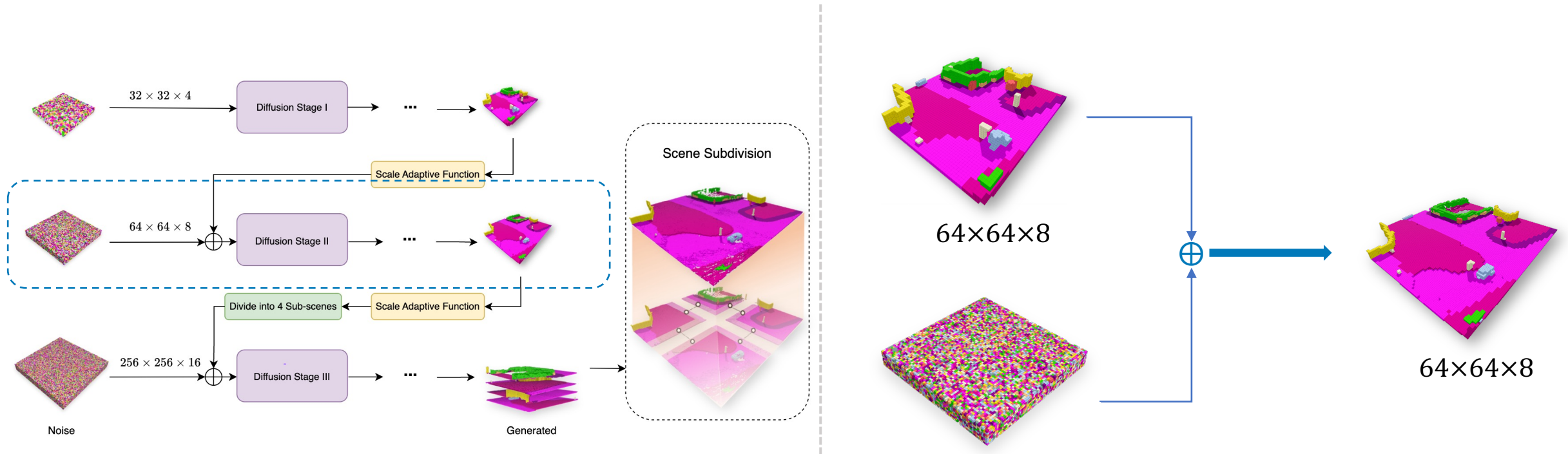
Stage II – Scale Adaptive Function



Upsampling by Scale Adaptive Function.

Method

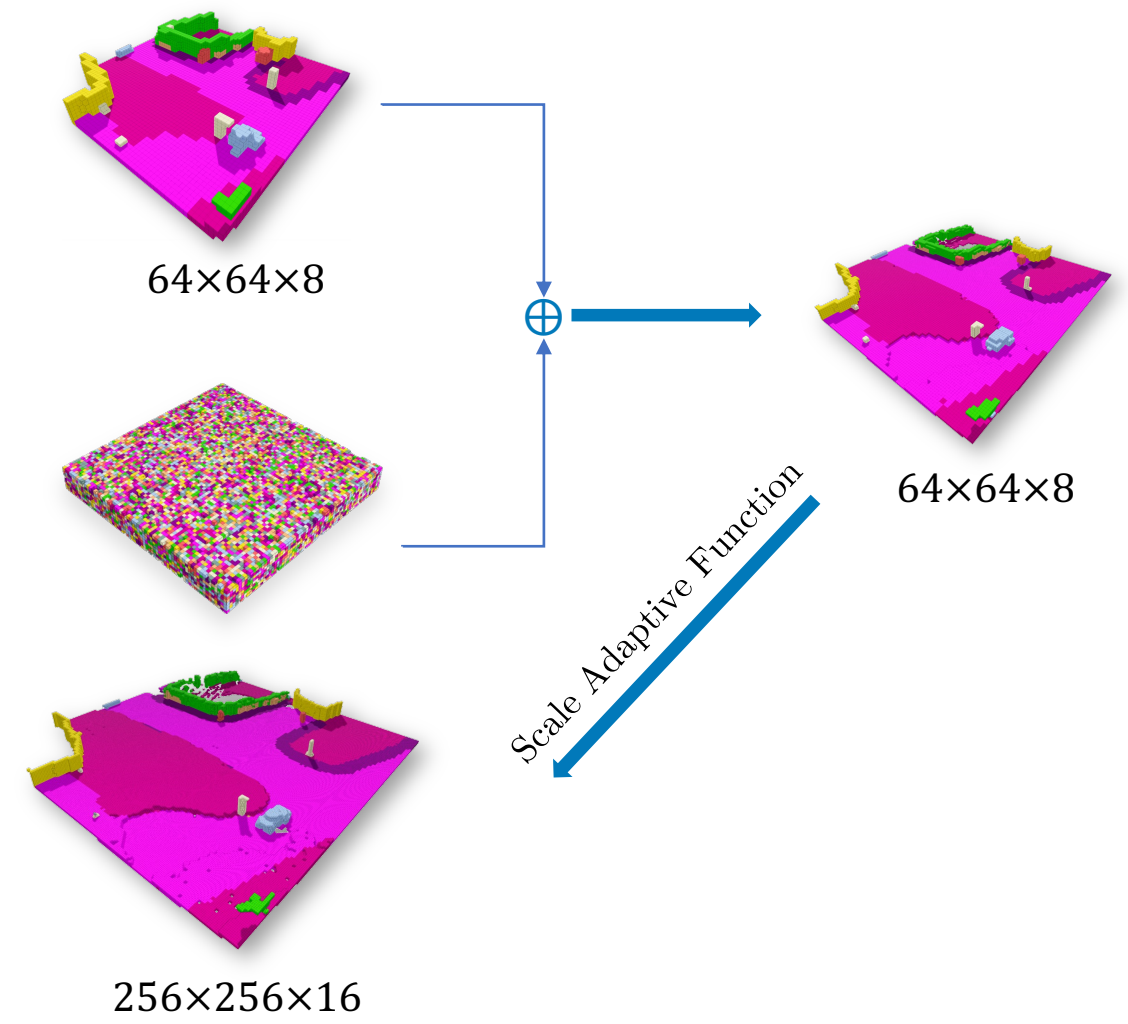
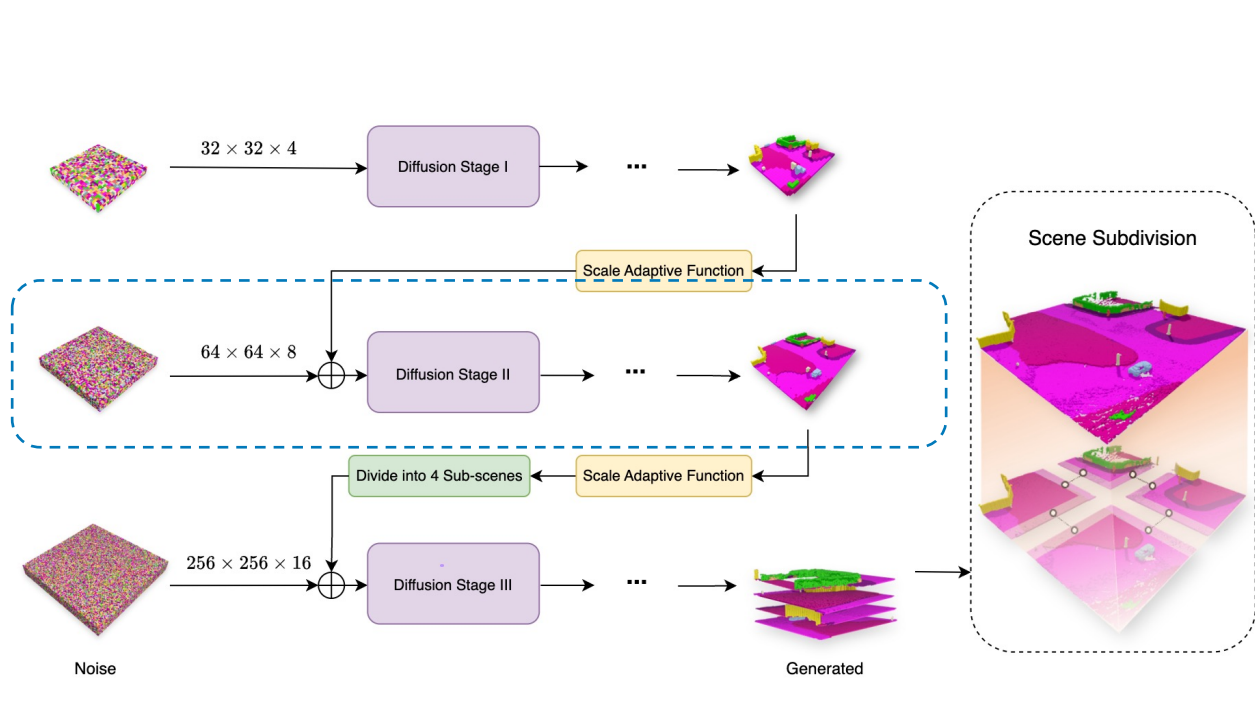
Stage II – Generation



Fine scenes generation conditioned on upsampled scenes.

Method

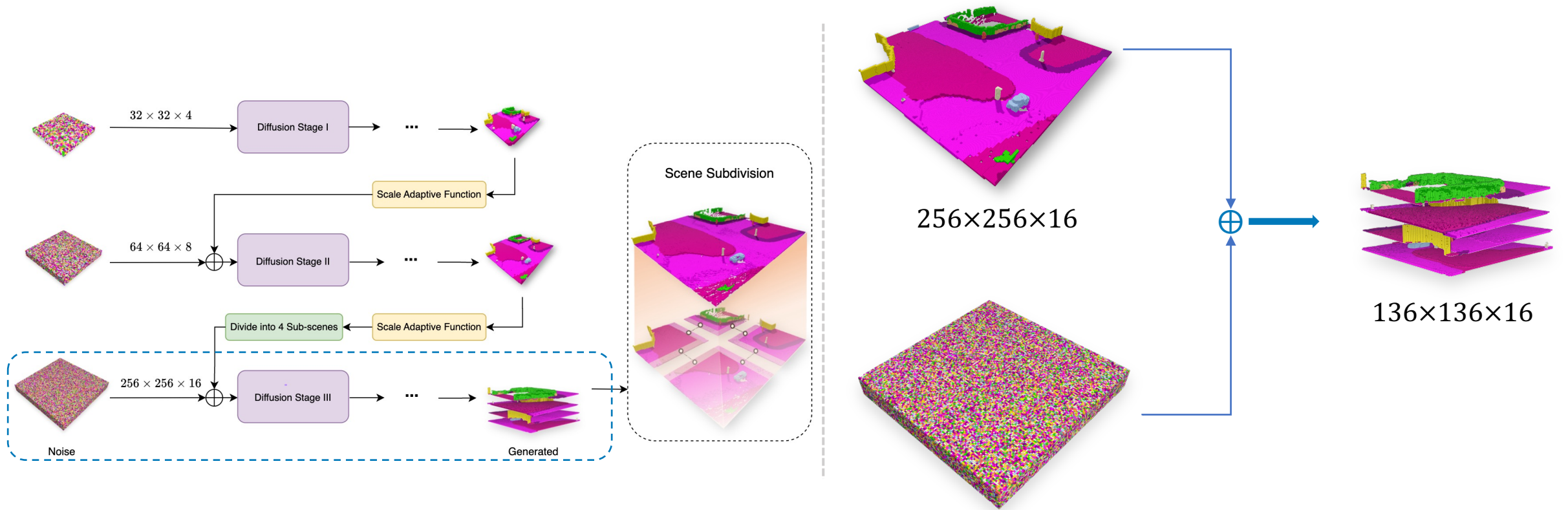
Stage II – Scale Adaptive Function



Upsampling by Scale Adaptive Function.

Method

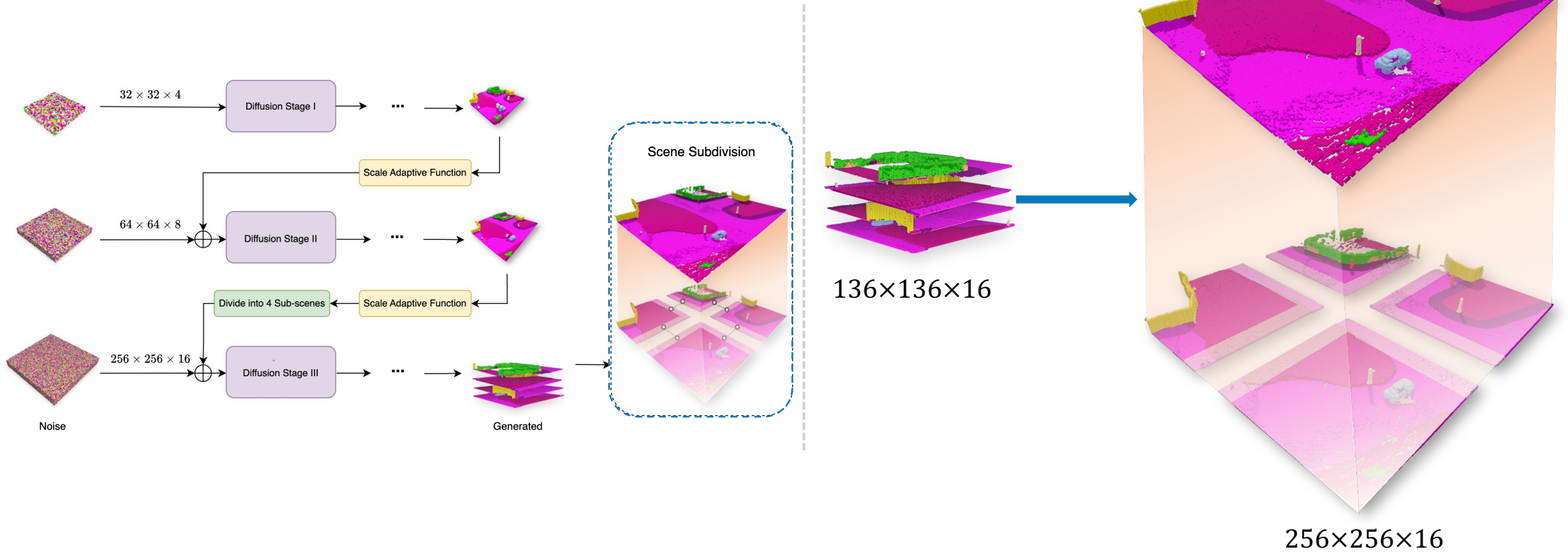
Stage III – Generation



Fine sub-scenes generation conditioned on upsampled scenes.

Method

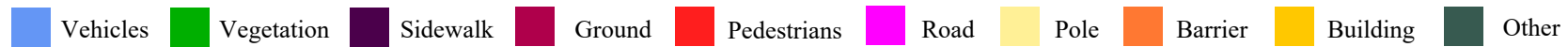
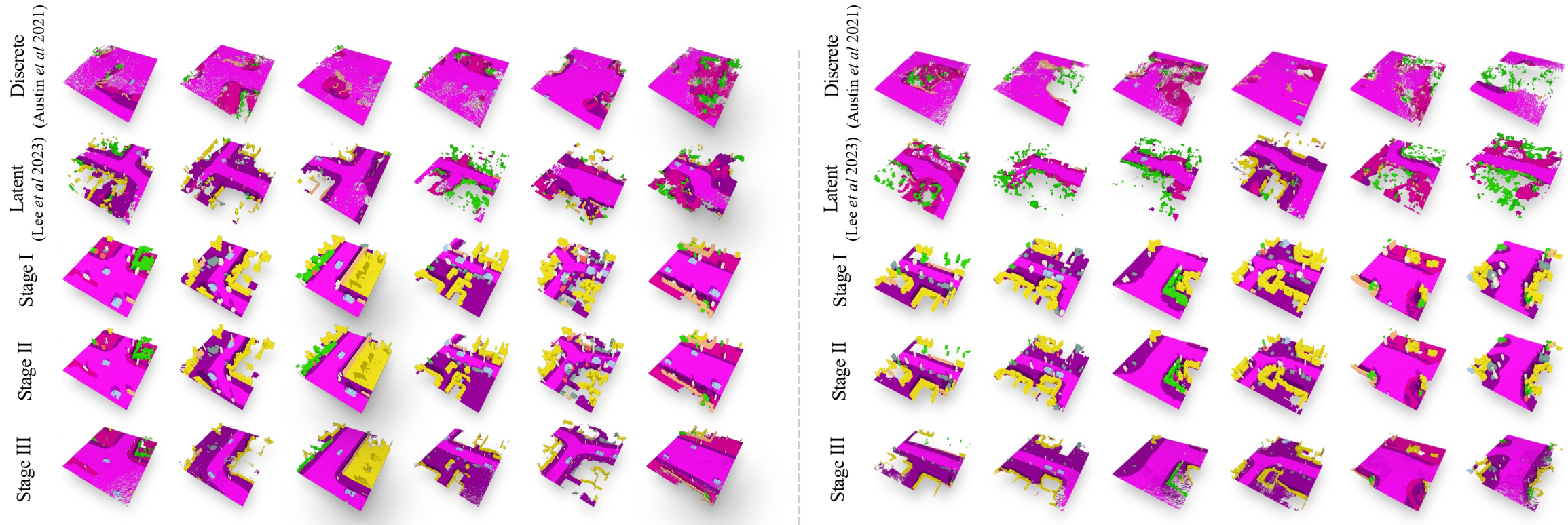
Stage III – Scene Subdivision and Merging



Composition of sub-scenes \rightarrow Final fine-grained scenes.

Qualitative Results

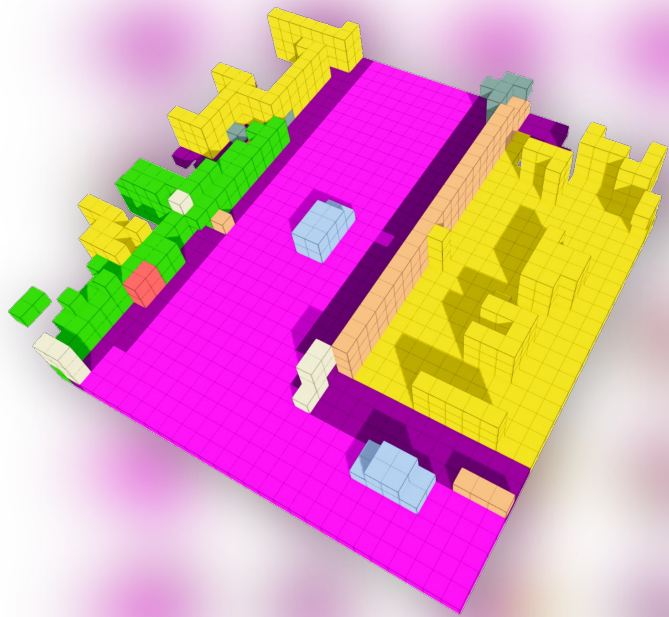
Unconditional Generation



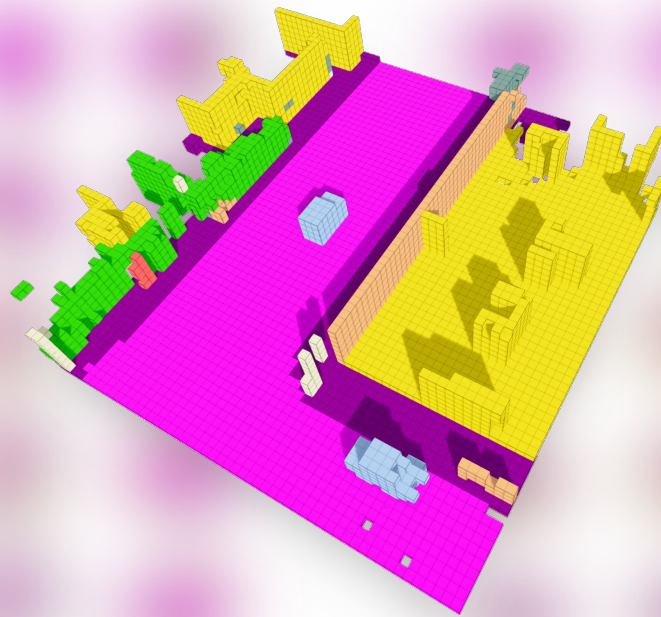
Qualitative Results

Unconditional Generation

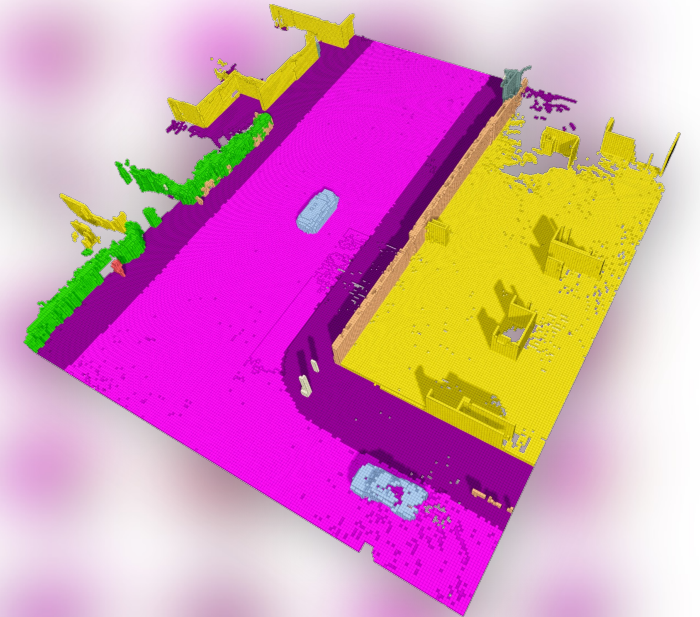
Our method can restore coarse scenes to finer scenes with high quality.



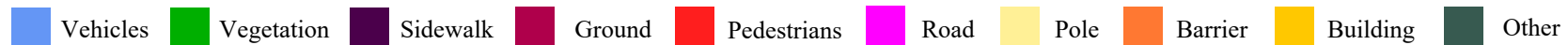
$32 \times 32 \times 4$



$64 \times 64 \times 8$

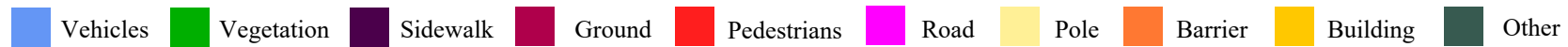
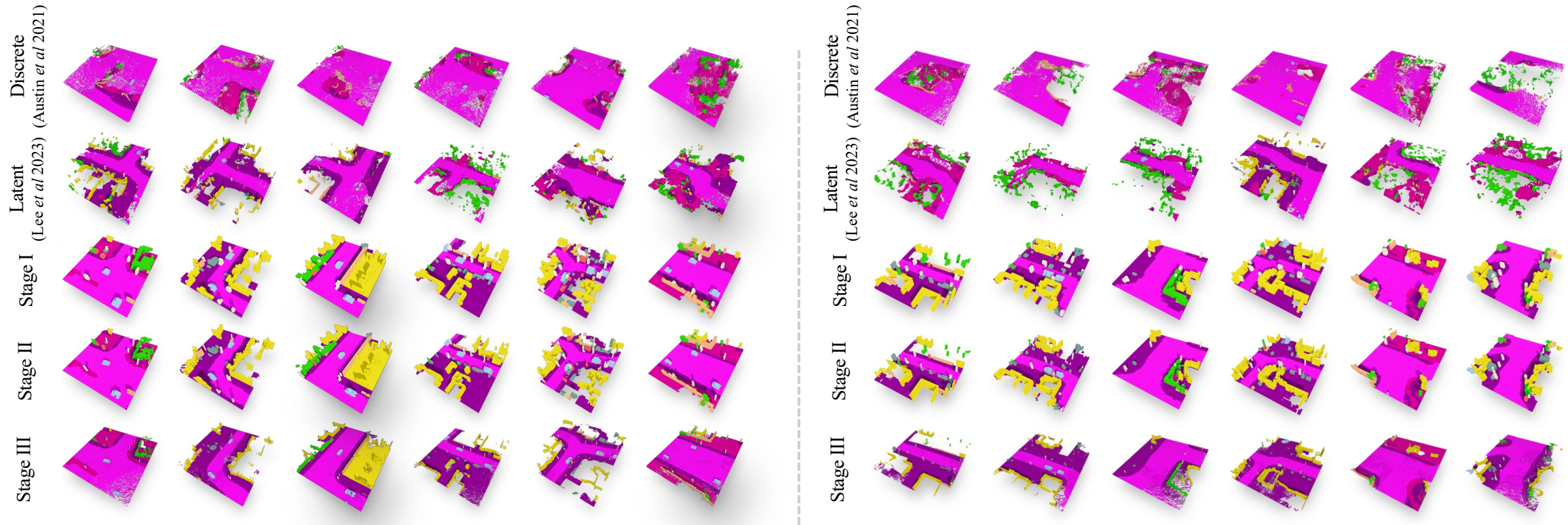


$256 \times 256 \times 16$



Qualitative Results

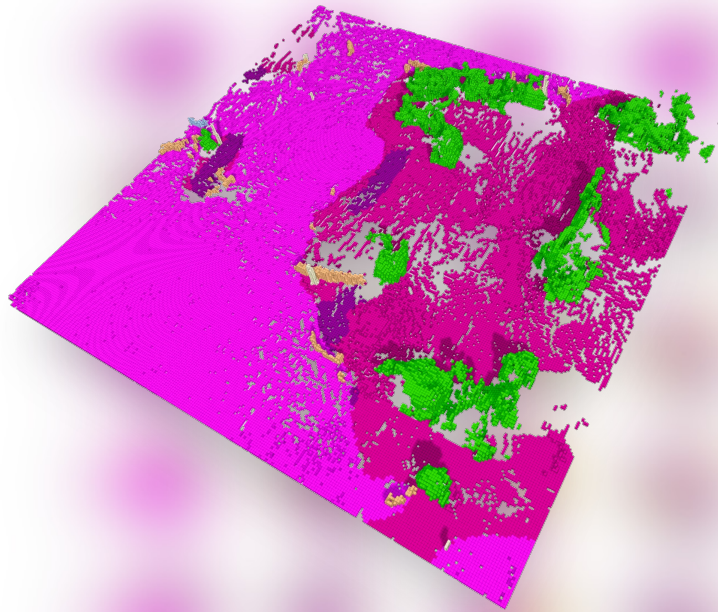
Unconditional Generation



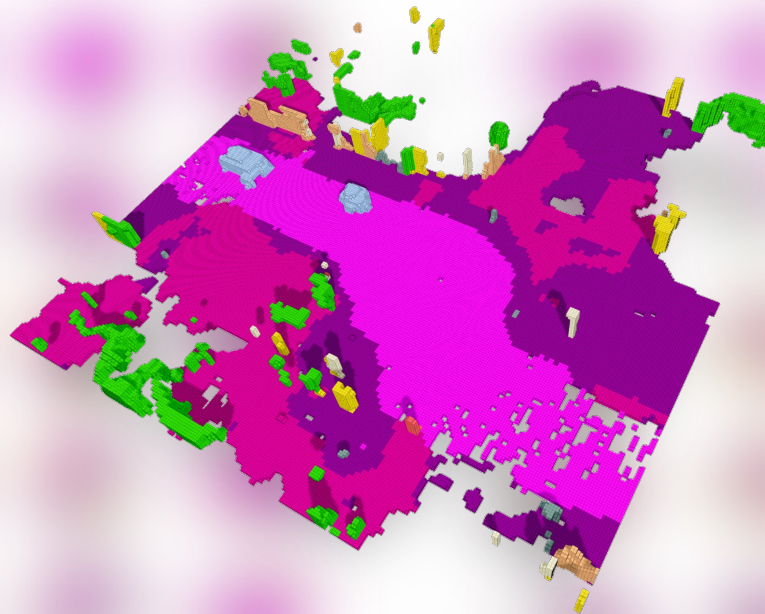
Qualitative Results

Unconditional Generation

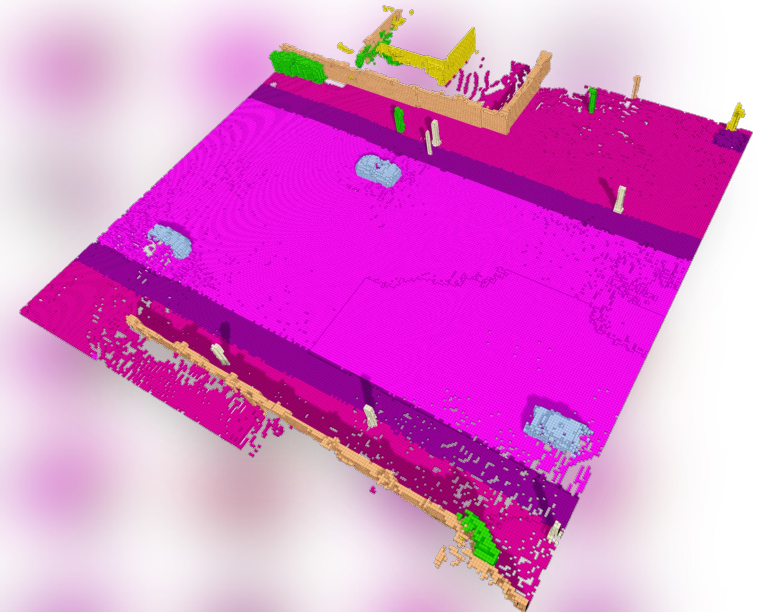
Our method can generate richer and more realistic scenes.



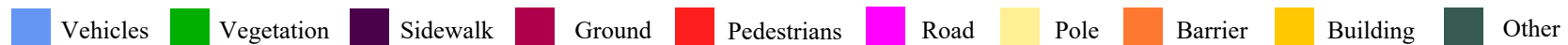
Discrete Diffusion
(Austin *et al* 2021)



Latent Diffusion
(Lee *et al* 2023)

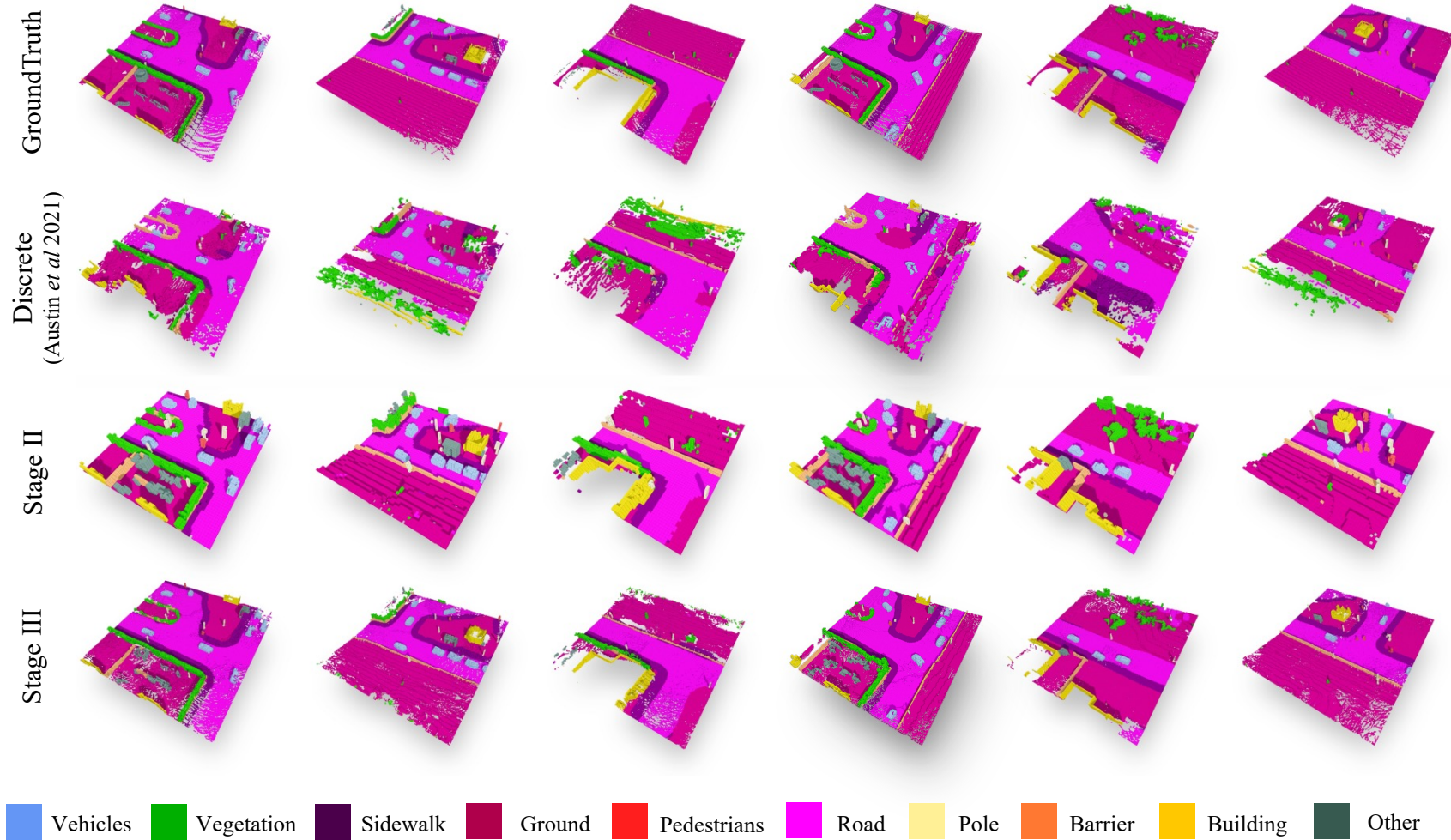


Ours



Qualitative Results

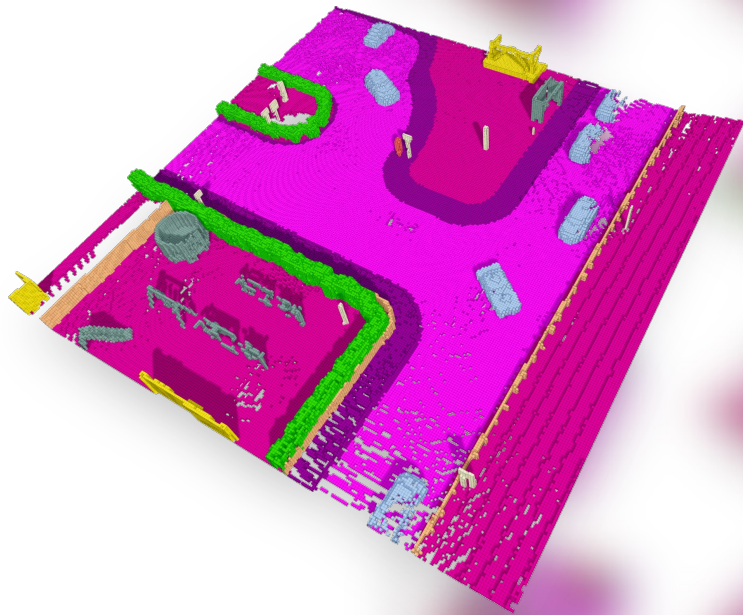
Conditional Generation



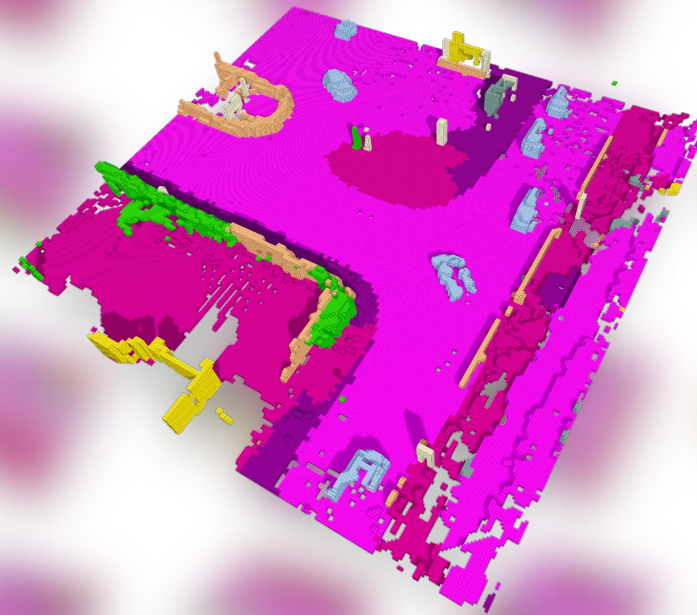
Qualitative Results

Conditional Generation

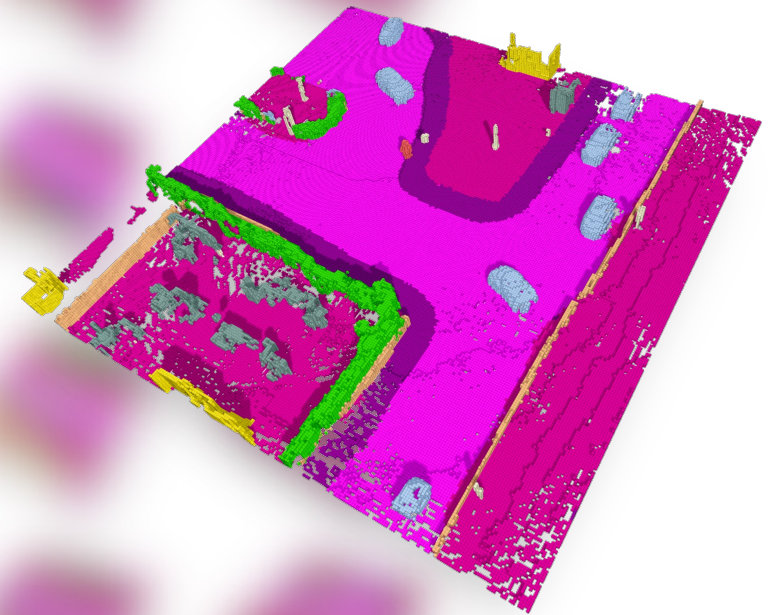
Conditions used by our method can restore scenes close to the ground truth.



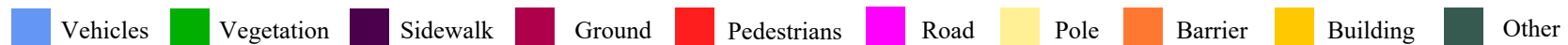
Ground Truth



(Austin *et al* 2021)
Condition: Point Cloud

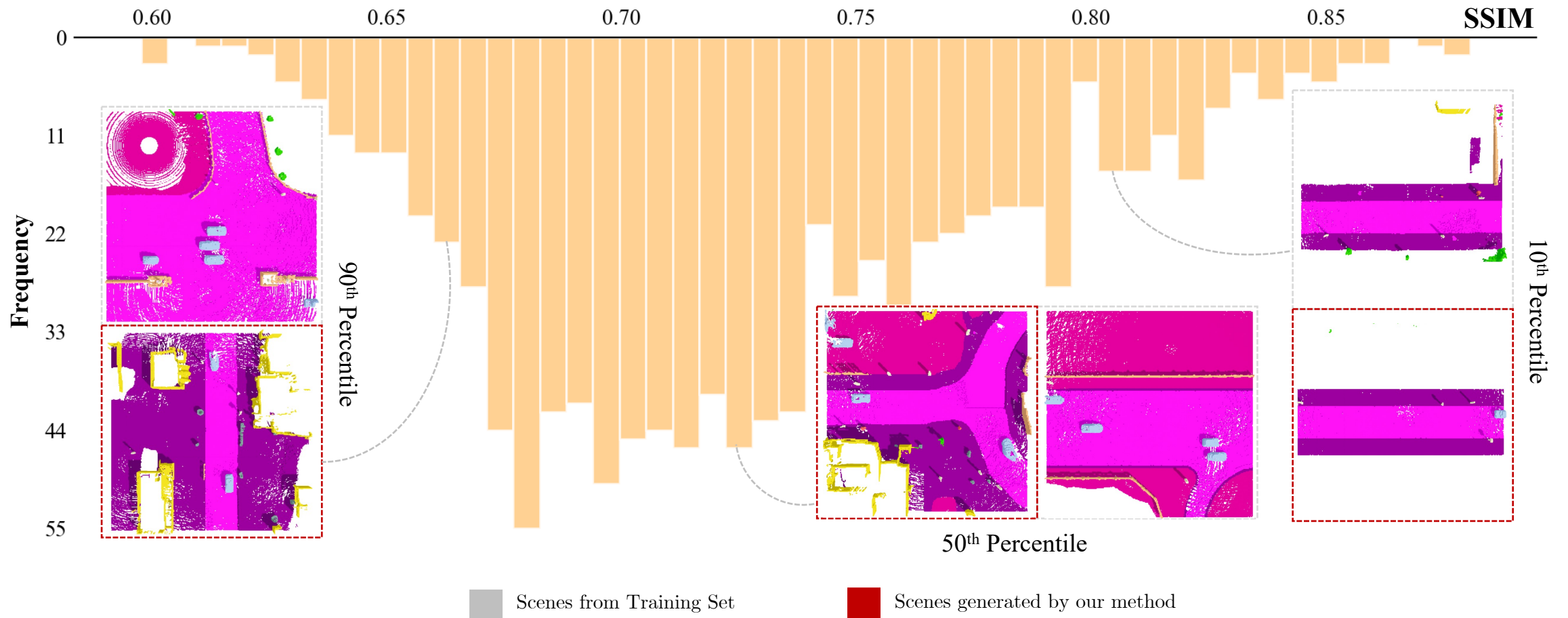


Ours
Condition: Prev Coarse Structure



Qualitative Results

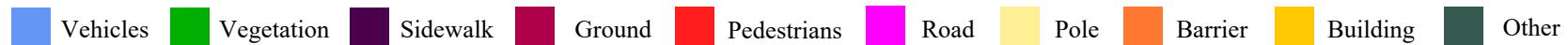
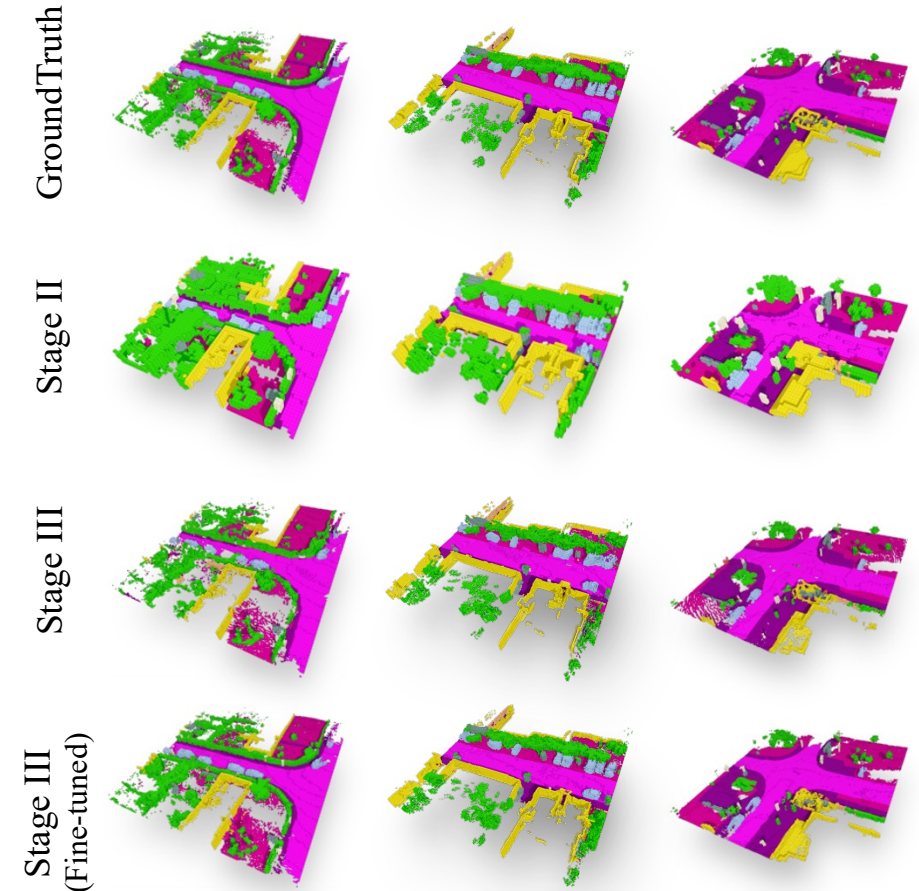
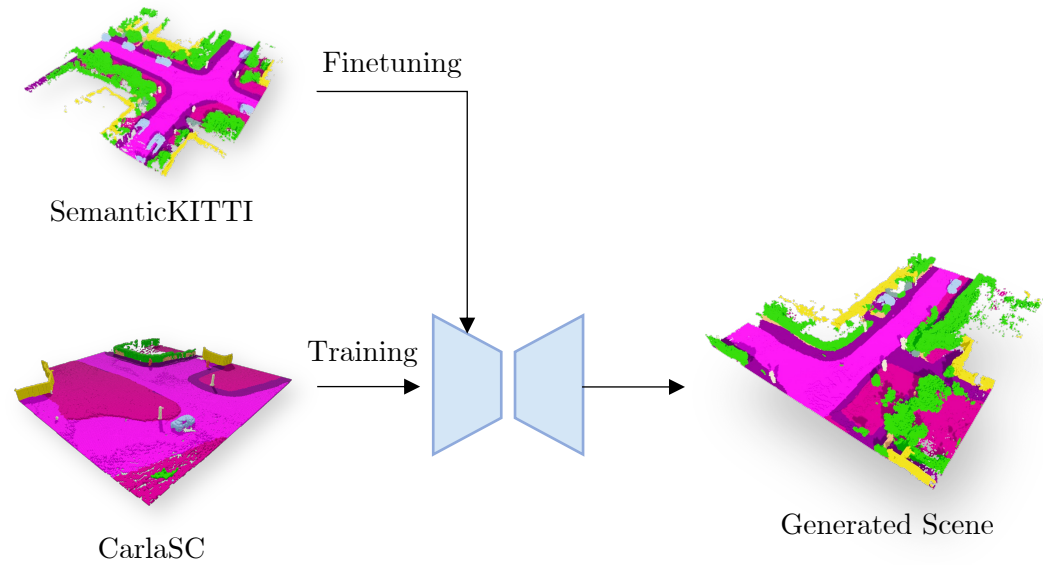
Our method does not simply memorize the scenes from the training set.



Applications

Cross-dataset Generation

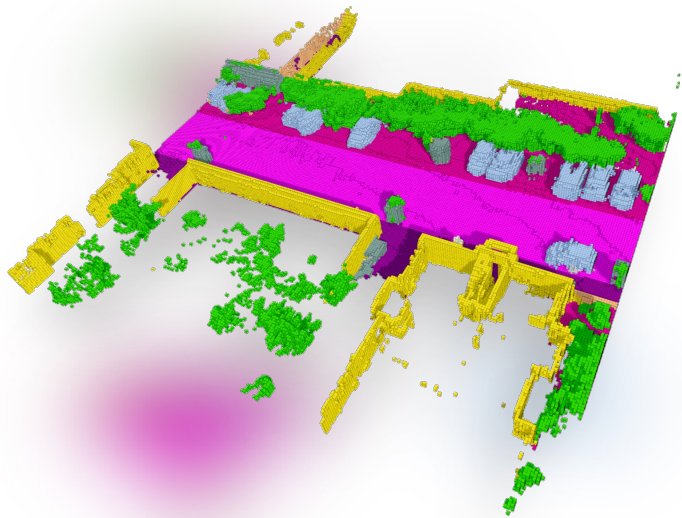
Fine-tune the model trained on the synthetic dataset using a small amount of real data.



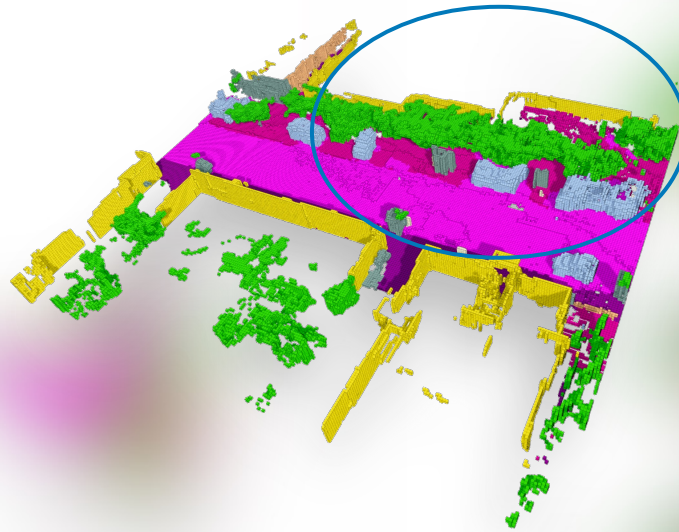
Applications

Cross-dataset Generation

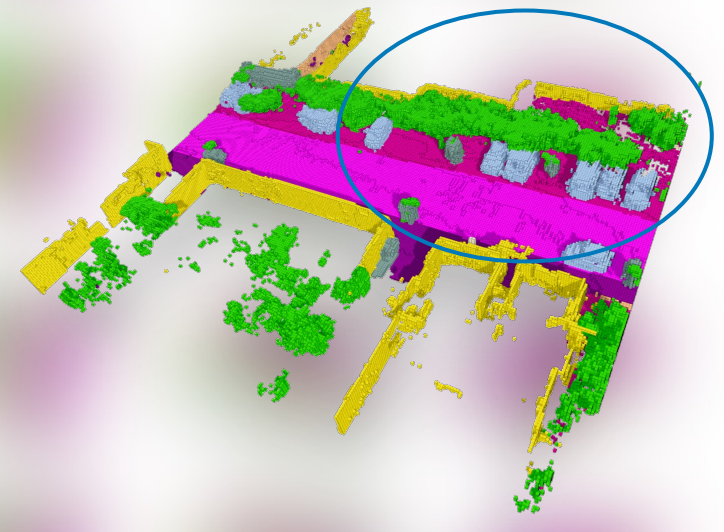
Fine-tuning with little data, our method gains better generative capabilities on real-world data.



Ground Truth



Ours
(trained on CarlaSC)



Ours
(fined-tuned on SemanticKITTI)



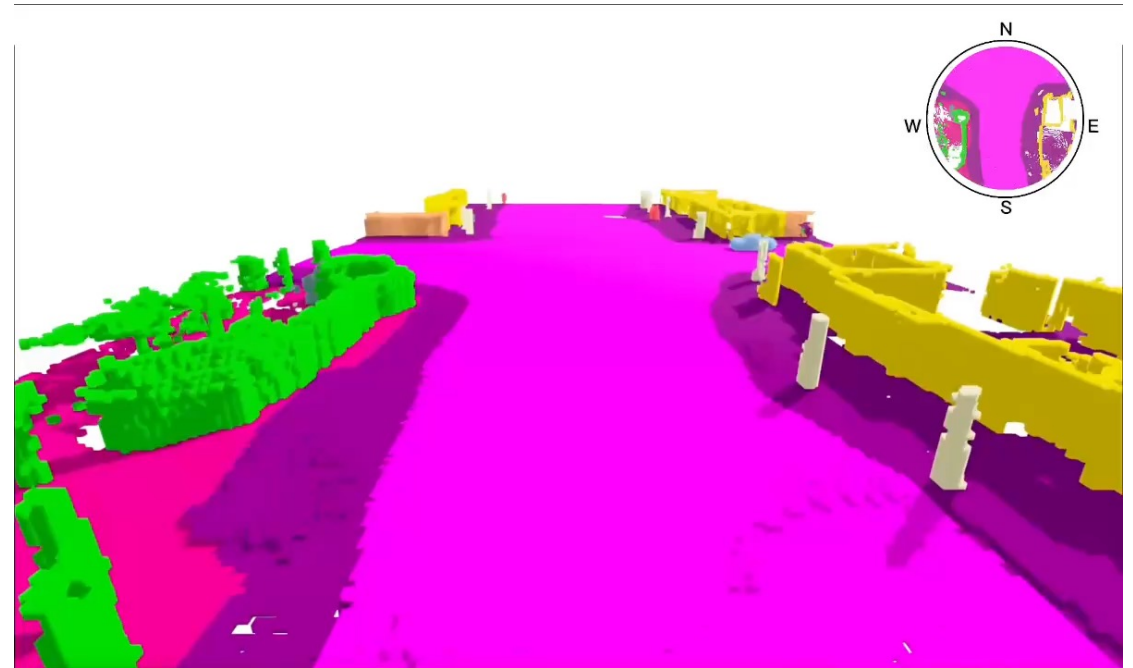
Applications

Infinite Scene Generation

Using the Scene Sub-division module approach, our method can generate infinite scenes.



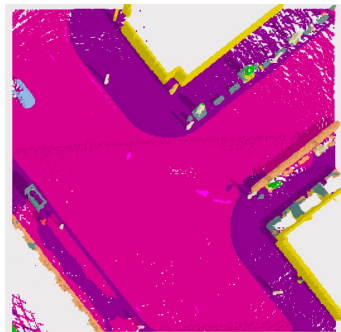
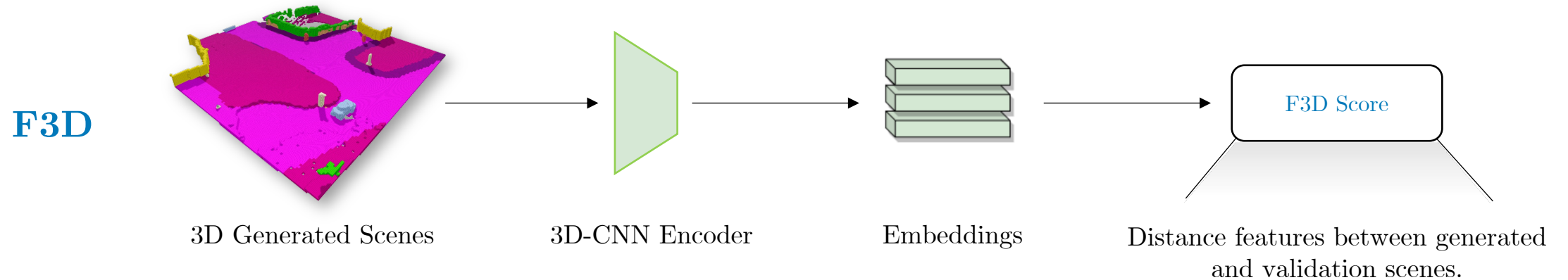
Method



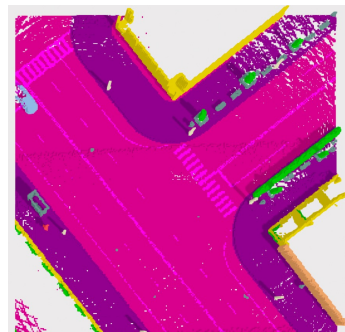
Demo Video: Generated Infinite Scene

Quantitative Evaluation

Evaluation Metrics



Predicted



Ground Truth

Semantic Segmentation

Gaussian Kernel

$$k(\mathbf{f}, \mathbf{f}') = \exp\left(-\frac{\|\mathbf{f} - \mathbf{f}'\|^2}{2\sigma^2}\right)$$

MMD Expression

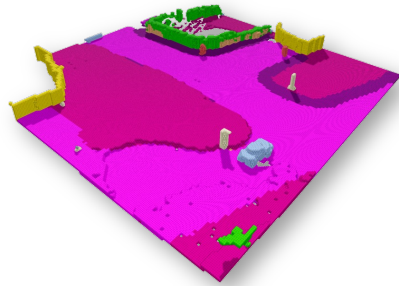
$$\text{MMD}^2 = \left\| \frac{1}{n} \sum_{i=1}^n k(\mathbf{f}_i, \cdot) - \frac{1}{m} \sum_{j=1}^m k(\mathbf{f}'_j, \cdot) \right\|^2$$

MMD

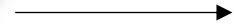
Quantitative Evaluation

Evaluation Metrics

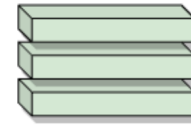
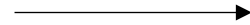
F3D



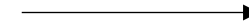
3D Generated Scenes



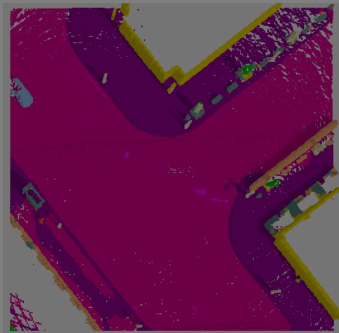
3D-CNN Encoder



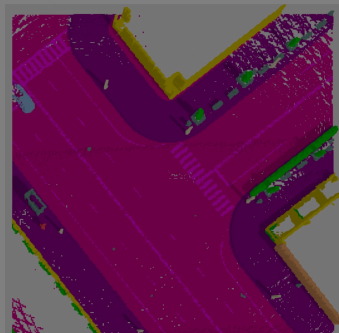
Embeddings



Distance features between generated and validation scenes.



Predicted



Ground Truth

Semantic Segmentation

Gaussian Kernel

$$k(\mathbf{f}, \mathbf{f}') = \exp\left(-\frac{\|\mathbf{f} - \mathbf{f}'\|^2}{2\sigma^2}\right)$$

MMD Expression

$$\text{MMD}^2 = \left\| \frac{1}{n} \sum_{i=1}^n k(\mathbf{f}_i, \cdot) - \frac{1}{m} \sum_{j=1}^m k(\mathbf{f}'_j, \cdot) \right\|^2$$

MMD

Quantitative Evaluation

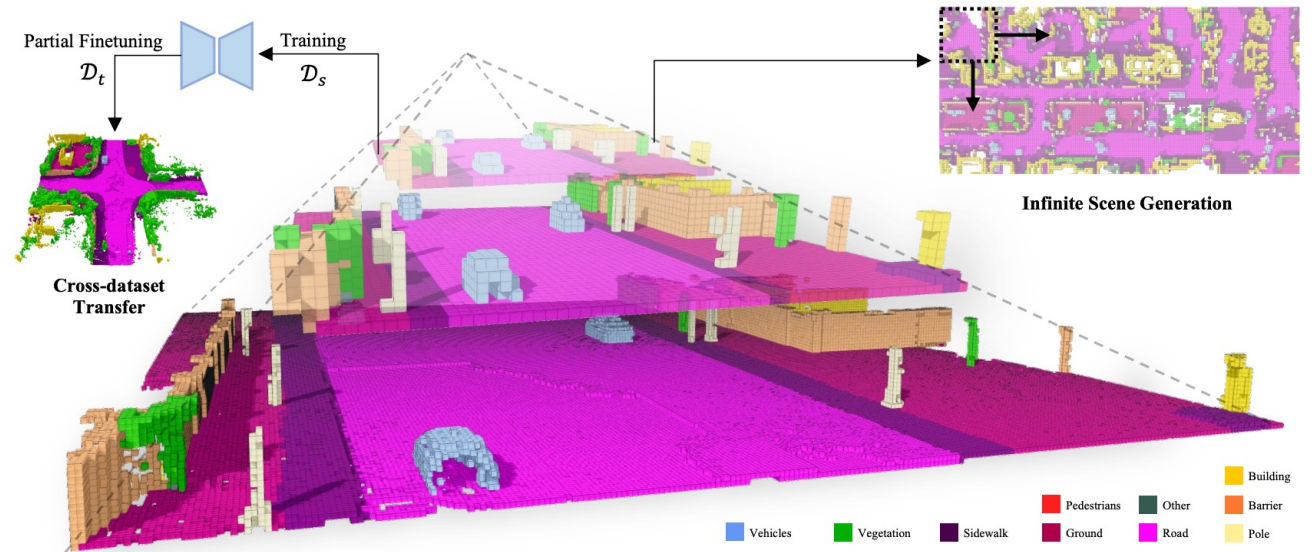
Generation Quality

Our method outperforms other methods on most metrics.

Method	Model	Condition	Segmentation Metric				Feature-based Metric	
			mIoU (V)	MA (V)	mIoU (P)	MA (P)	F3D (\downarrow)	MMD (\downarrow)
Ground Truth	-	-	52.19	72.40	32.90	47.68	0.246	0.108
Unconditioned	DiscreteDiff	-	40.05	63.65	25.54	38.71	1.361	0.599
	LatentDiff	-	38.01	62.39	26.69	45.87	0.331	0.221
	P-DiscreteDiff (Ours)	-	68.02	85.66	33.89	52.12	0.315	0.200
Conditioned	DiscreteDiff	Point Cloud	38.55	59.97	28.41	44.06	0.357	0.261
	DiscreteDiff	Coarse scene (s_1)	52.52	77.23	27.93	43.13	0.359	0.284
	P-DiscreteDiff (Ours)	Coarse scene (s_1)	55.75	78.70	29.78	46.61	0.342	0.274

Conclusion

- Generate high-quality scenes with decent computational resources.
- Introduce metrics for evaluating the quality of 3D scene generation.
- Showcase two applications: cross-dataset learning and infinite scene generation.



Pyramid Diffusion for Fine 3D Large Scene Generation

Poster Session: Oct 3 (today), 4:30pm – 6:30pm, #6 Session, #158 Board

Project Page & Code

