CityDreamer

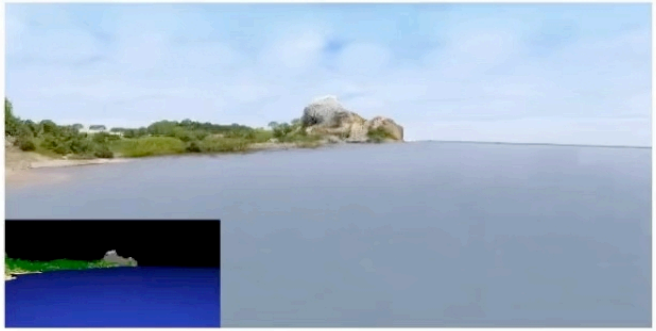
Compositional Generative Model of Unbounded 3D Cities

Authors: Haozhe Xie et al.
Affiliation: Nanyang Technological University
Conference: CVPR 2024

Presenter: Zhiguo Liu
Date: November 19

Georgia State University

Challenges: Previous Works



GANCraft [CVPR'21]

SceneDreamer [arXiv 2303.01330]

InfiniCity [ICCV'23]

Challenges: Natural Data vs City Data

Similarity

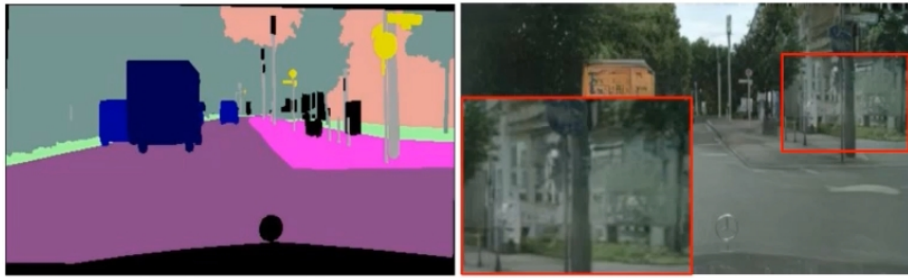
· Objects are similar in natural scenes

· Buildings are diverse in cities

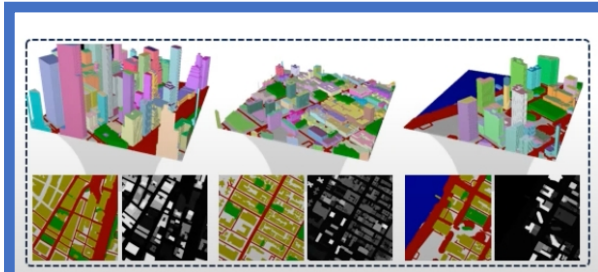
Realistic

· Synthetic nature images are realistic

· Synthetic city images are not realistic

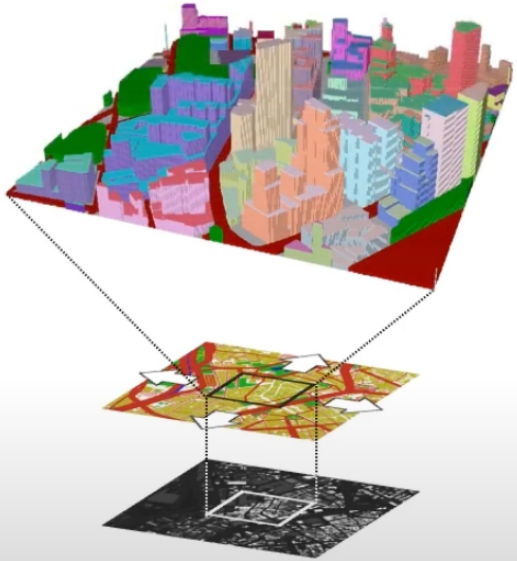


The Proposed Method: Datasets



(b) OSM Dataset: Real-world City Layout

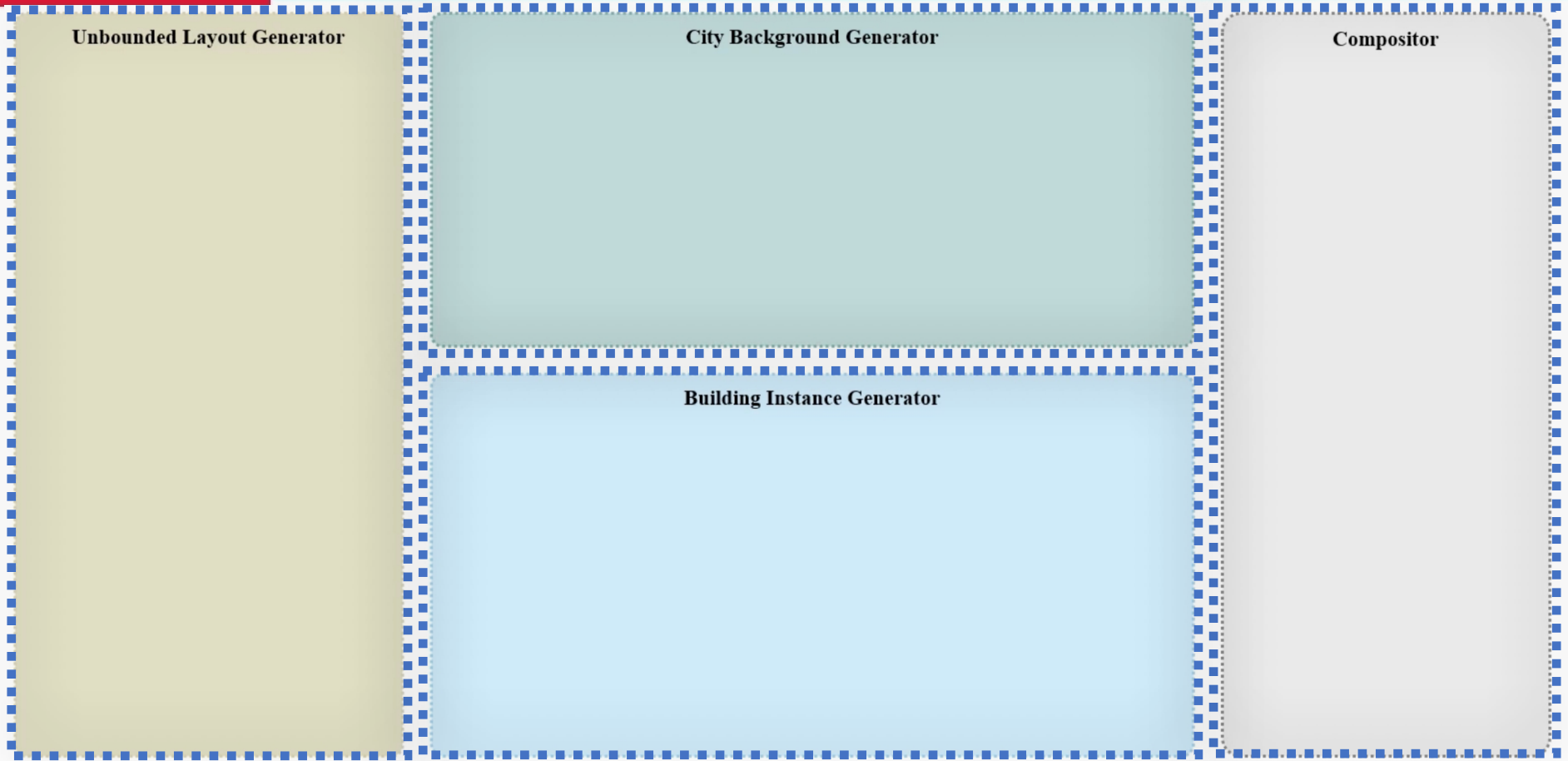
(a) Google Earth Dataset: Real-world City Appearance



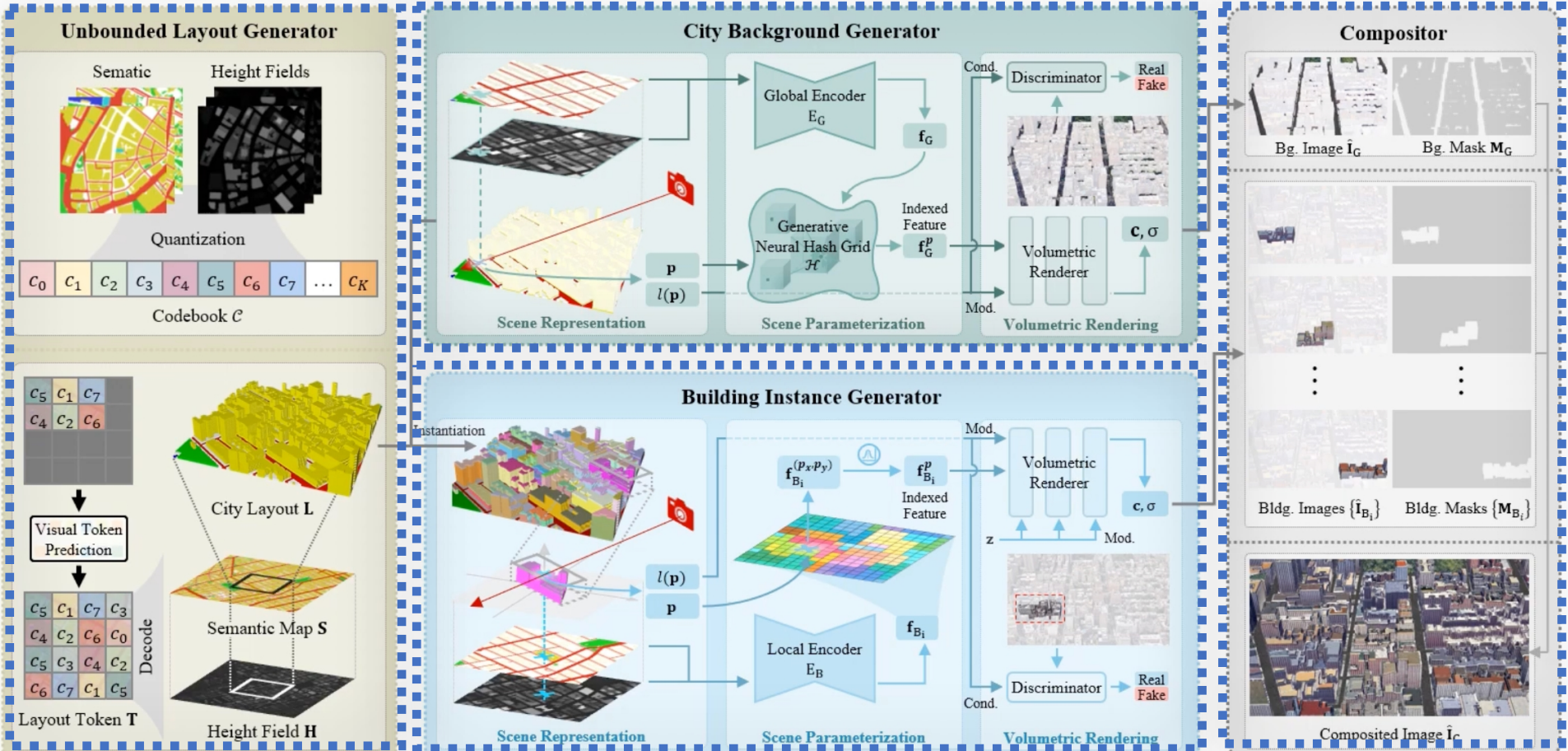
(c) Unbounded City Layout Generation

(d) CityDreamer Generated 3D Cities

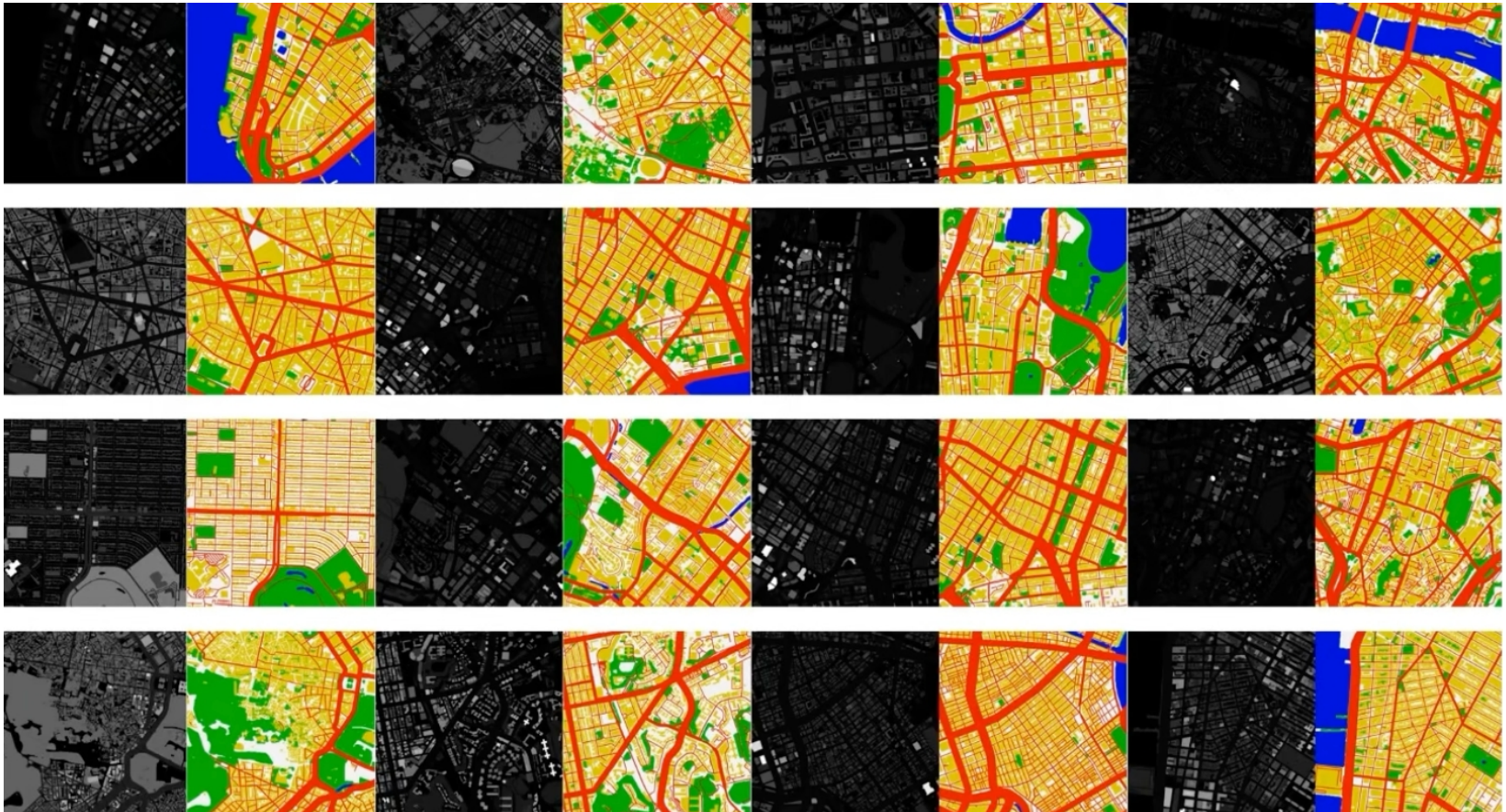
The Proposed Method: Structure



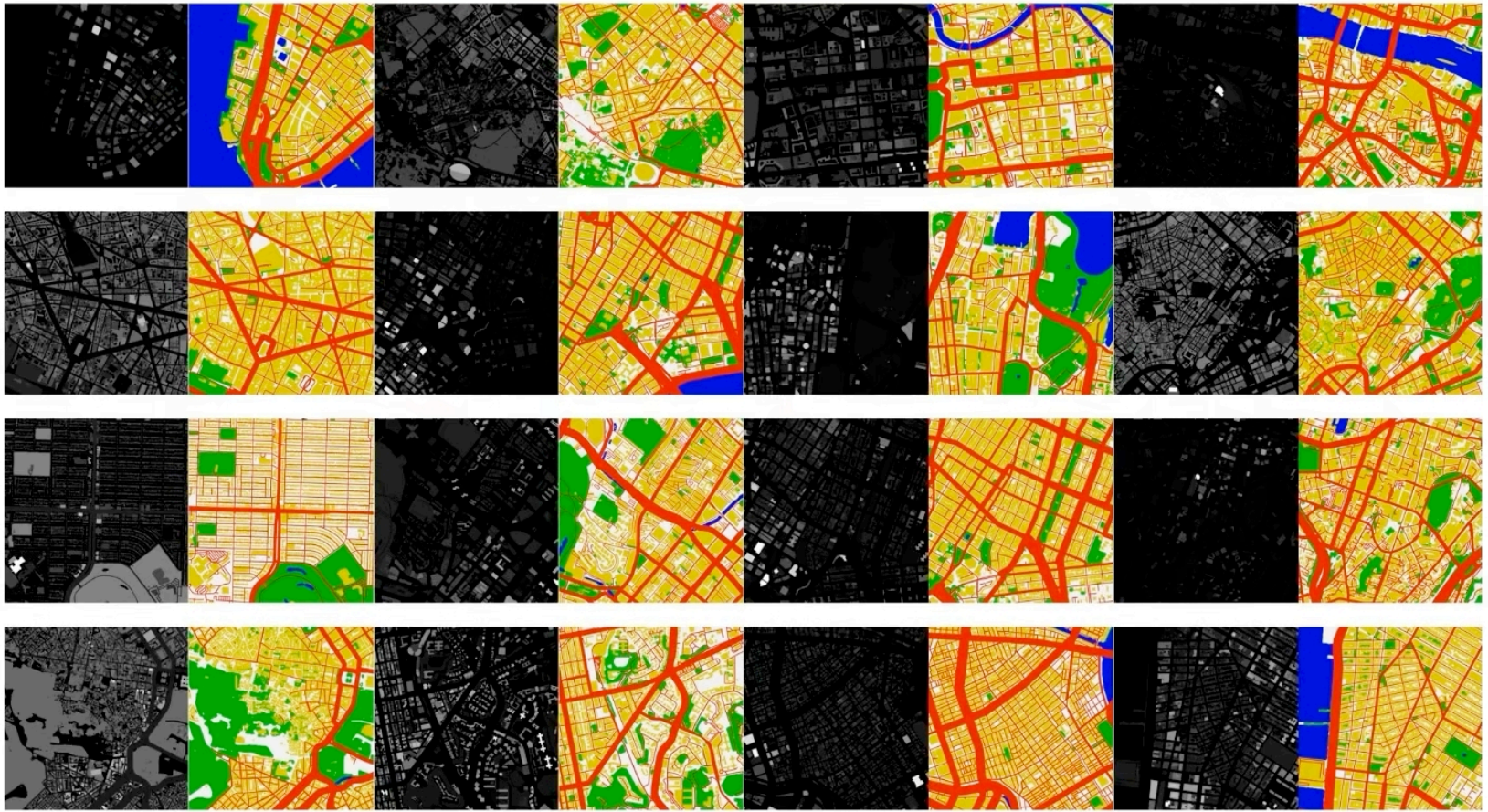
The Proposed Method: Structure



The OSM Dateset



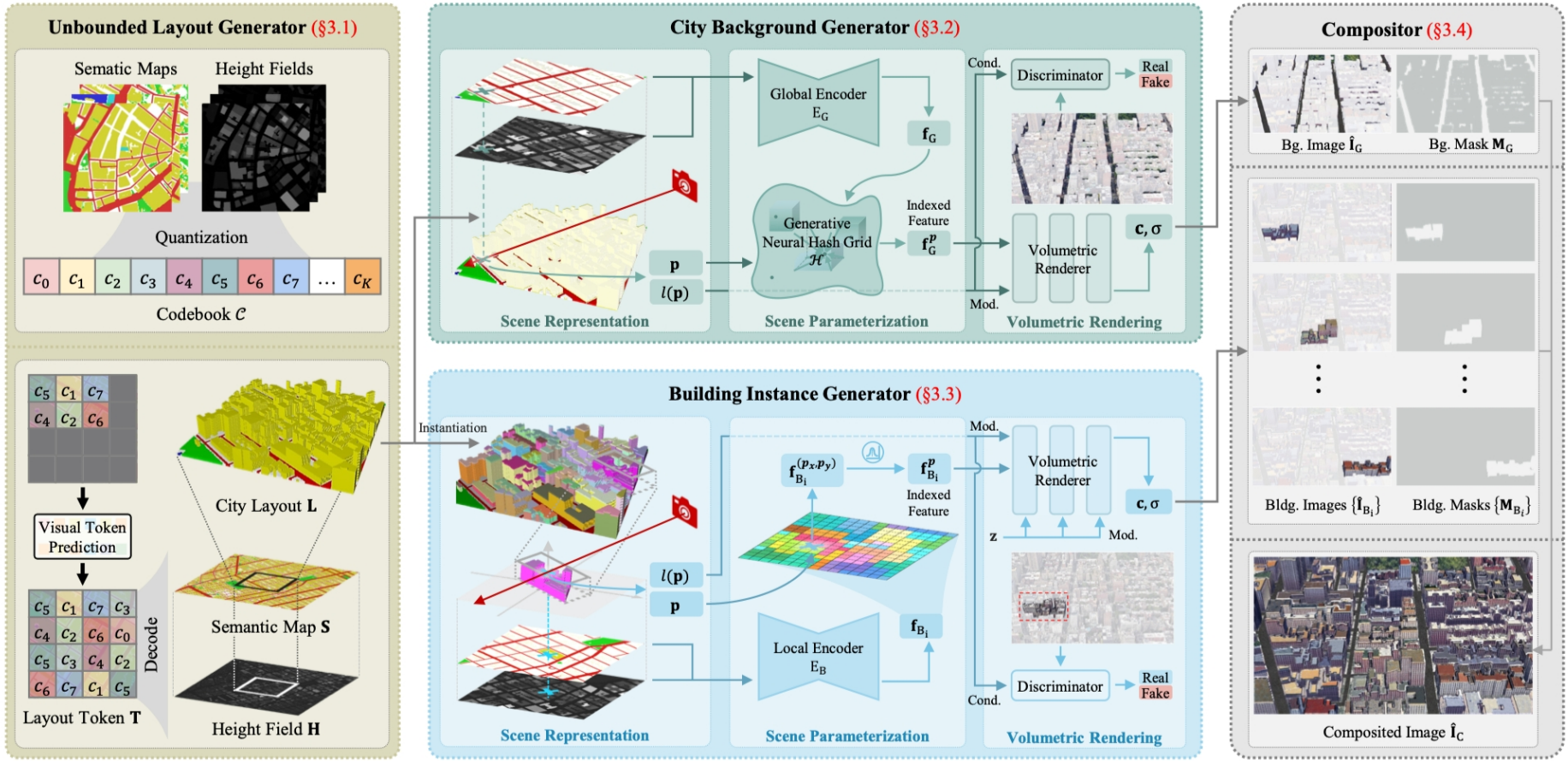
The OSM Dateset



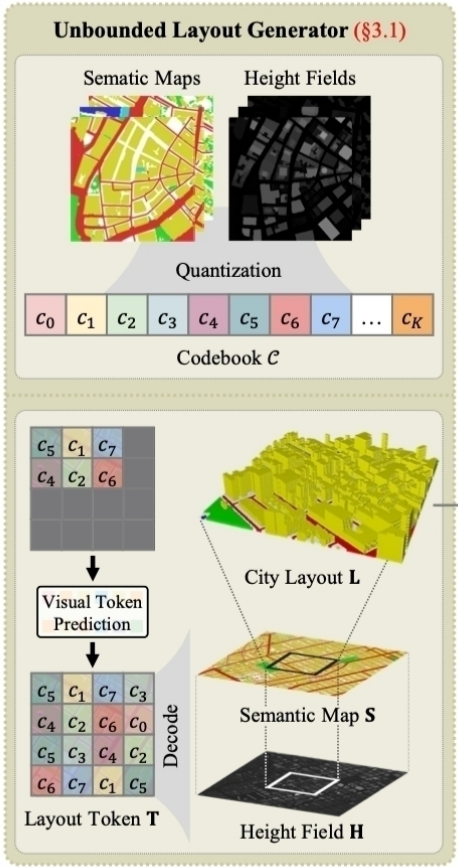
The GoogleEarth Dateset

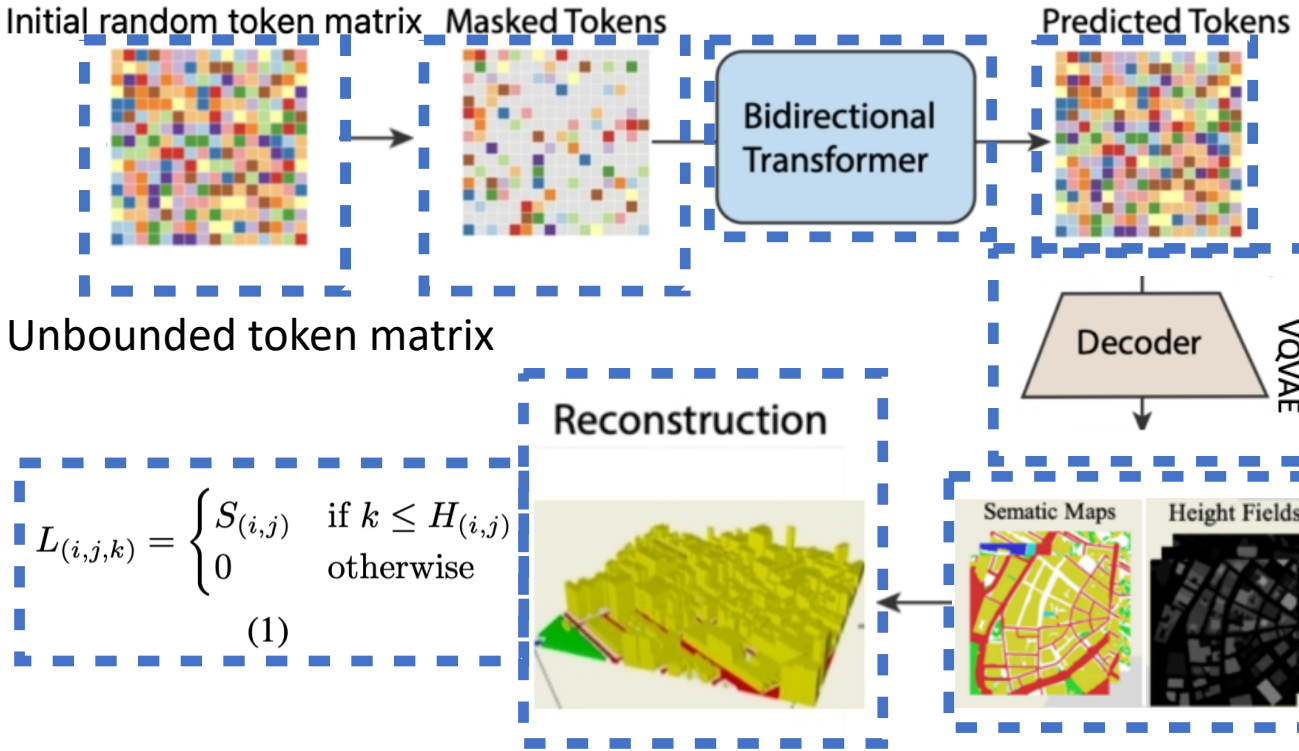
Comparison to SOTA Methods

Details


Unbounded Layout Generator (§3.1)
Sematic Maps
Height Fields
Quantization
Codebook $\mathcal{C}$
Visual Token Prediction
Layout Token T
Decode
City Layout L
Semantic Map S
Height Field H
City Background Generator (§3.2)
Global Encoder E_G
Generative Neural Hash Grid $\mathcal{H}$
Indexed Feature
Scene Representation
Scene Parameterization
Volumetric Rendering
Cond.
Discriminator
Real
Fake
Mod.
Volumetric Renderer
Building Instance Generator (§3.3)
Instantiation
Local Encoder E_B
Compositor (§3.4)
Bg. Image $\hat{\mathbf{I}}_G$
Bg. Mask $\mathbf{M}_G$
Bldg. Images $\{\hat{\mathbf{I}}_{B_i}\}$
Bldg. Masks $\{\mathbf{M}_{B_i}\}$
Composited Image $\hat{\mathbf{I}}_C$

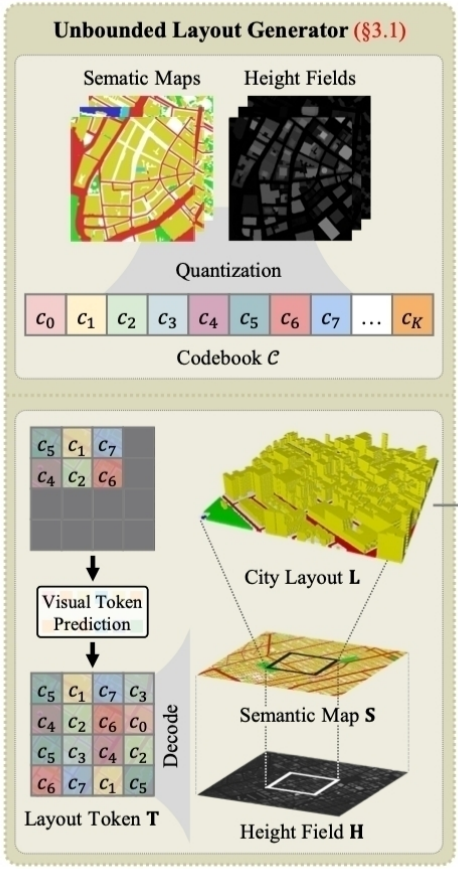
Unbouded Layout Generator: Maskgit Step

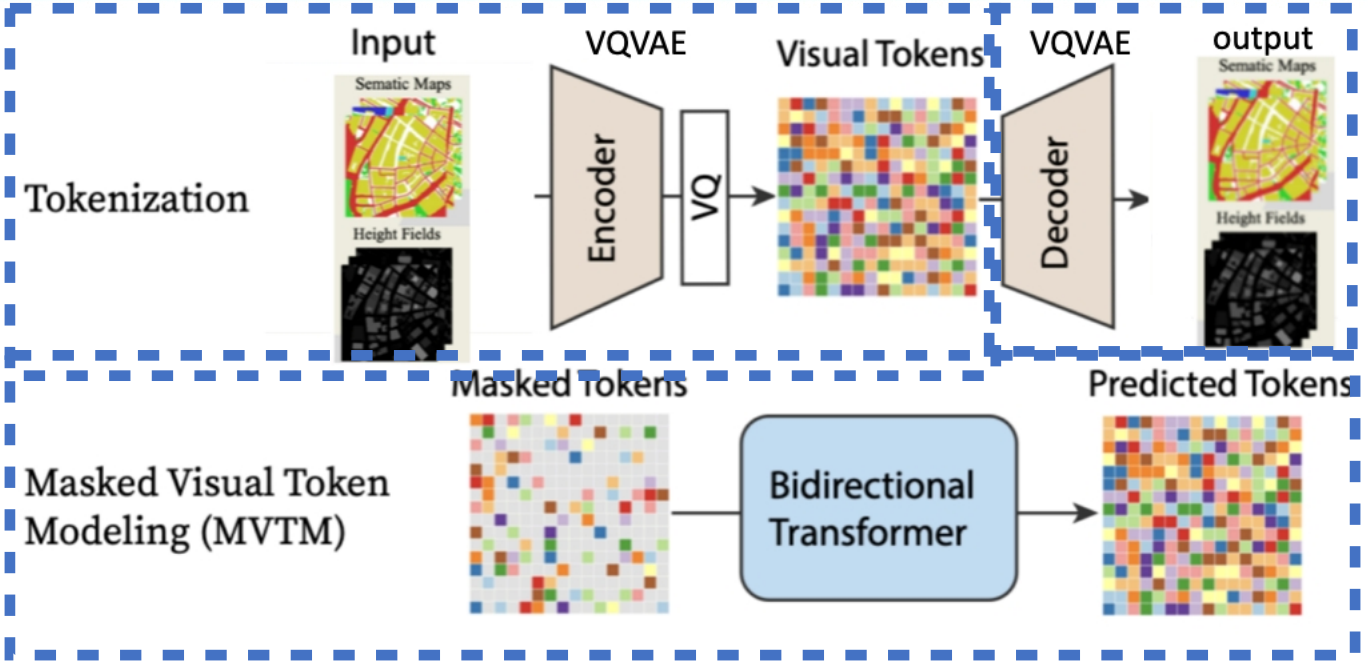




L: Layout , S: semantic map, H: height field map, i,j,k: 3D position

Unbouded Layout Generator: Maskgit Model

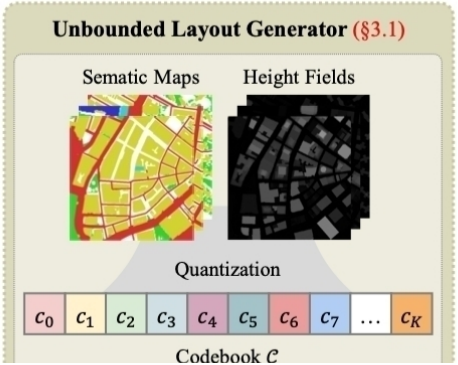




$$Reweighted\_ELBO = w_1 ELBO + w_2 D_{KL}$$

ELBO: Evidence Lower Bound loss, D_{KL}: Kullback-Leibler Divergence w_1, w_2: weight parameter

Unbouded Layout Generator—VQVAE

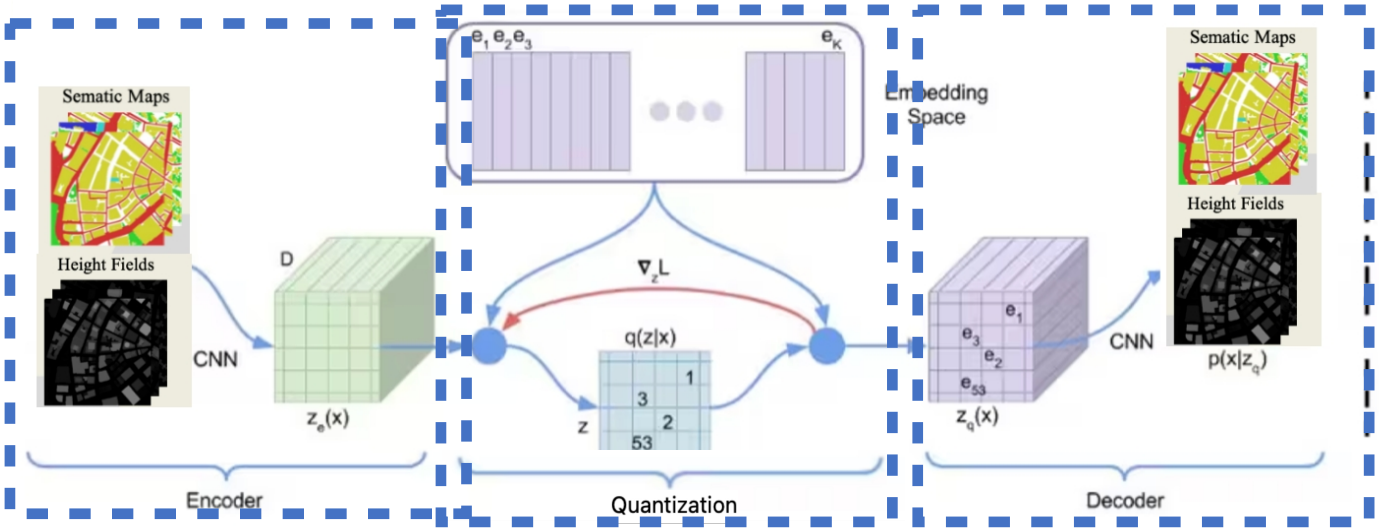


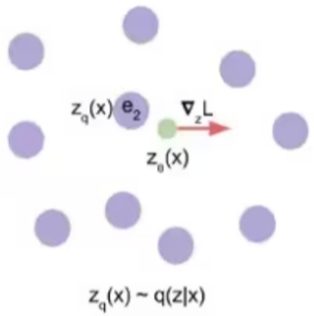
$$Loss = L_1Loss + SmoothnessLossS + CrossEntropyLossE$$

$$\ell_{\mathrm{VQ}} = \lambda_{\mathrm{R}}\|\hat{\mathbf{H}}_{\mathbf{p}} - \mathbf{H}_p\| + \lambda_{\mathrm{S}}\mathcal{S}(\hat{\mathbf{H}}_{\mathbf{p}}, \mathbf{H}_p) + \lambda_{\mathrm{E}}\mathcal{E}(\hat{\mathbf{S}}_{\mathbf{p}}, \mathbf{S}_p) \quad (2)$$

ℓ_{VQ}: VQVAE loss function, $\lambda_{\mathrm{R}}, \lambda_{\mathrm{S}}, \lambda_{\mathrm{E}}$: weight parameter

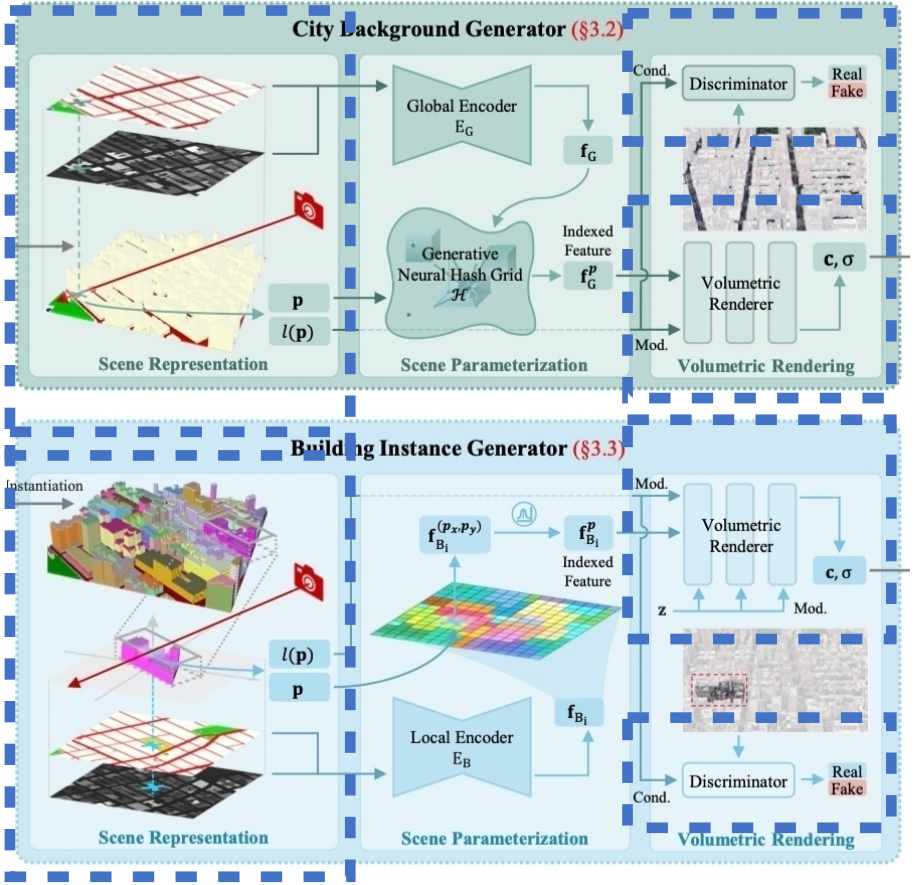
$\hat{\mathbf{H}}_{\mathbf{p}}, \hat{\mathbf{S}}_{\mathbf{p}}$: Generated Height field and semantic map patches, $\mathbf{H}_{\mathbf{p}}, \mathbf{S}_{\mathbf{p}}$: corresponding ground truth.



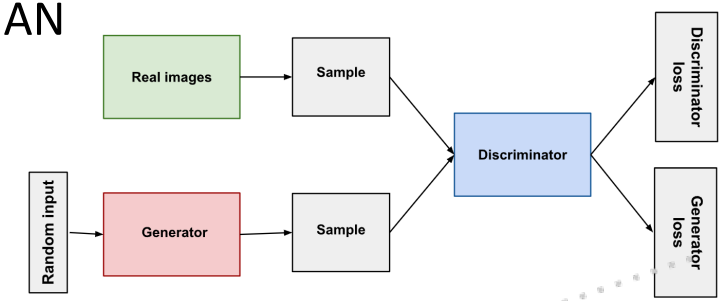


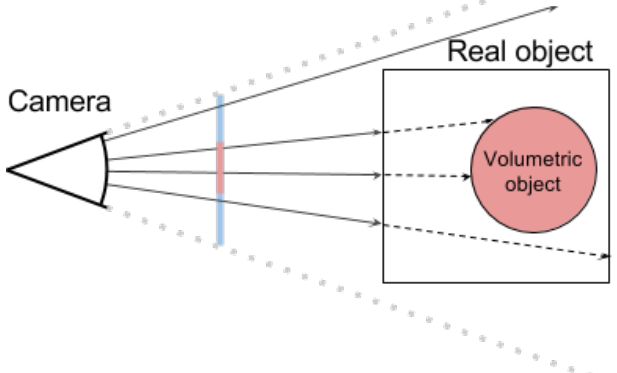
codebook

City Background vs Building Generator: Same

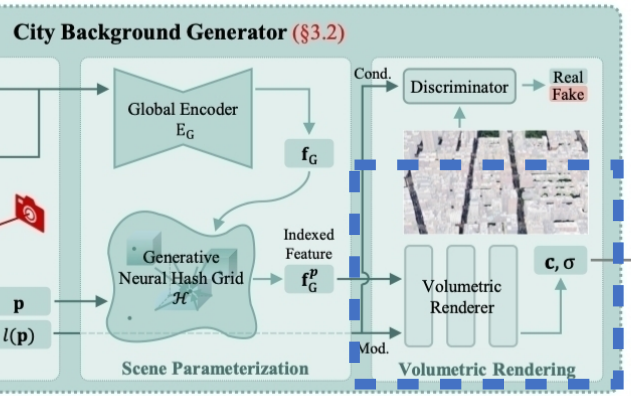


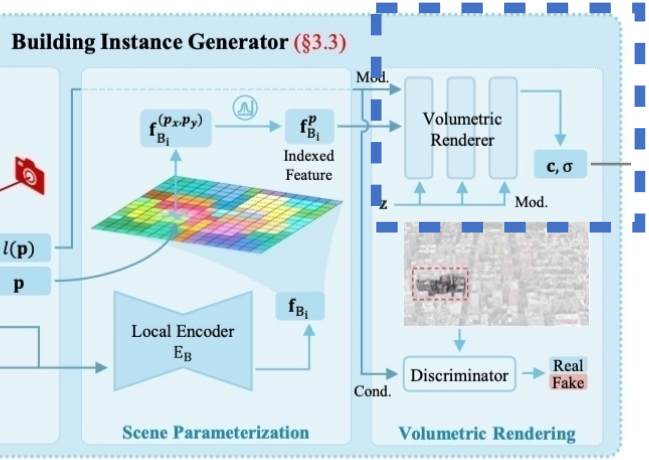
Same
- Bird'd Eye View
- Volumetric Renderer
- GAN



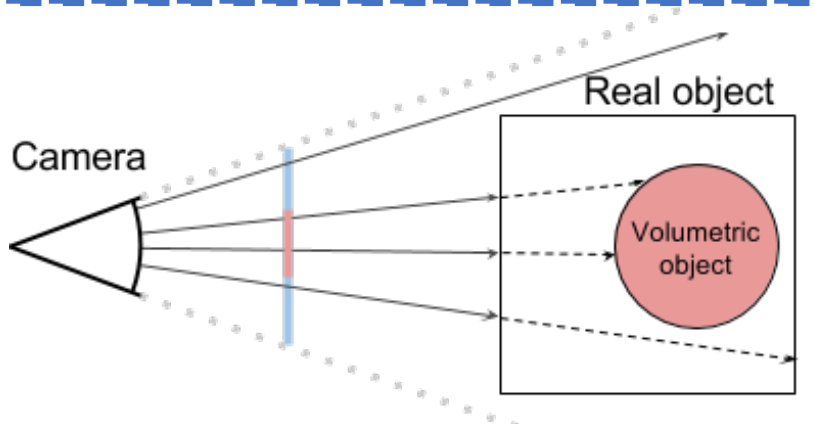


City Background vs Building Generator: Volumetric Renderer





$$C(\mathbf{r}) = \int_0^\infty T(t)\mathbf{c}(\mathbf{f}_{\mathrm{G}}^{\mathbf{r}(t)}, l(\mathbf{r}(t)))\boldsymbol{\sigma}(\mathbf{f}_{\mathrm{G}}^{\mathbf{r}(t)})dt$$

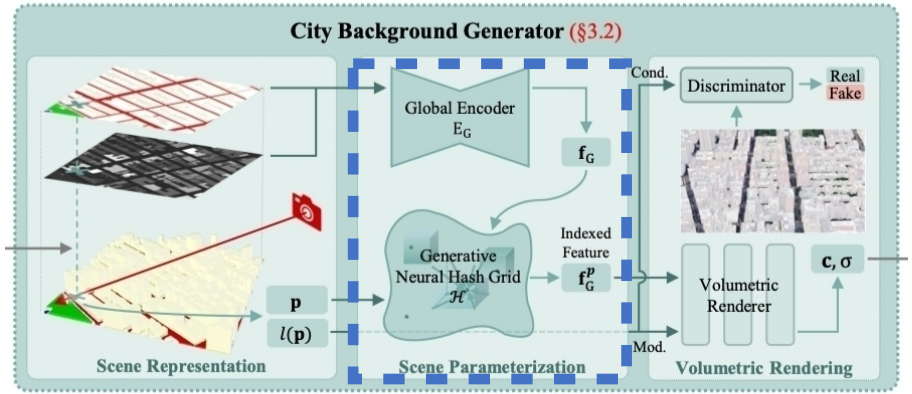


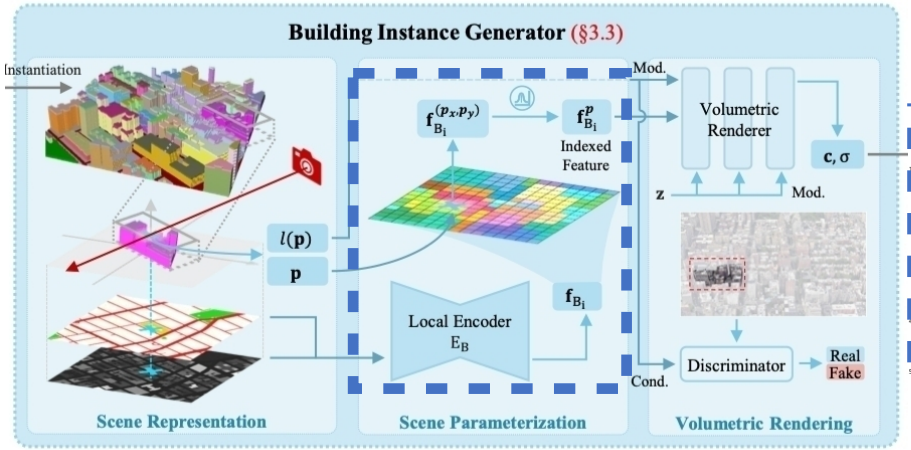
$$C(\mathbf{r}) = \int_0^\infty T(t)\mathbf{c}(\mathbf{f}_{\mathrm{B_i}}^{\mathbf{r}(t)}, \mathbf{z}, l(\mathbf{r}(t)))\boldsymbol{\sigma}(\mathbf{f}_{\mathrm{B_i}}^{\mathbf{r}(t)})dt$$

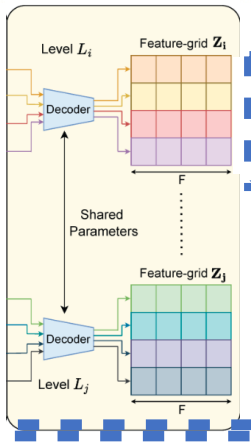
$C(\mathbf{r})$: pixel value, $T(t)$: transmittance, $\mathbf{c}()$: color value, $\mathbf{f}_{\mathrm{B_i}}^{\mathbf{r}(t)}$: volumetric grid feature,

$\mathbf{z}$: style code, $l()$: ray direction, $\mathbf{r}(t)$: ray path, $\boldsymbol{\sigma}()$: volume density,

City Background & Building Generator: diffrent encode strategy







$$\mathbf{f}_{\mathrm{G}}^{\mathbf{p}} = \mathcal{H}(\mathbf{p}, \mathbf{f}_{\mathrm{G}}) = \left(\bigoplus_{i=1}^{d_G} f_{\mathrm{G}}^{i} \pi^{i} \bigoplus_{j=1}^{3} p^{j} \pi^{j} \right) \mod T$$

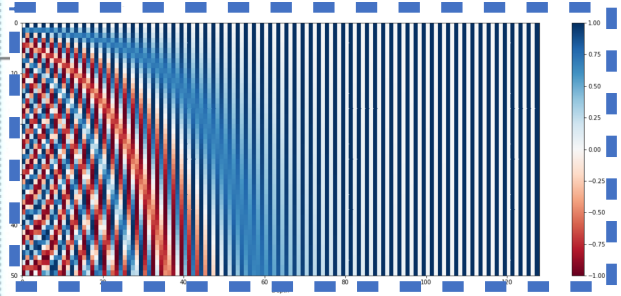
Generative Neural Hash Grid

f_G^P: feature index position

p: 3D position, f_G: global feature vector

π^i, π^j: random integer constant

$$\mathcal{O}(x) = \{\sin(2^i \pi x), \cos(2^i \pi x)\}_{i=0}^{N_P^L - 1} \quad (9)$$

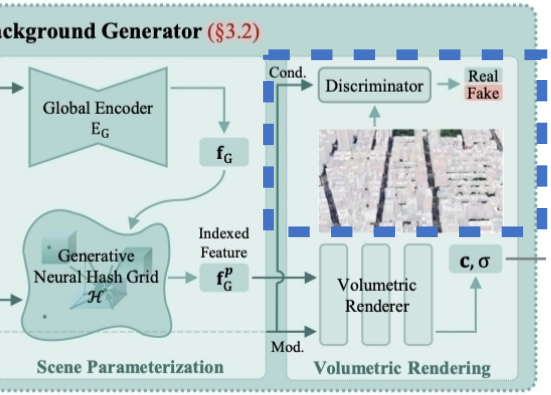


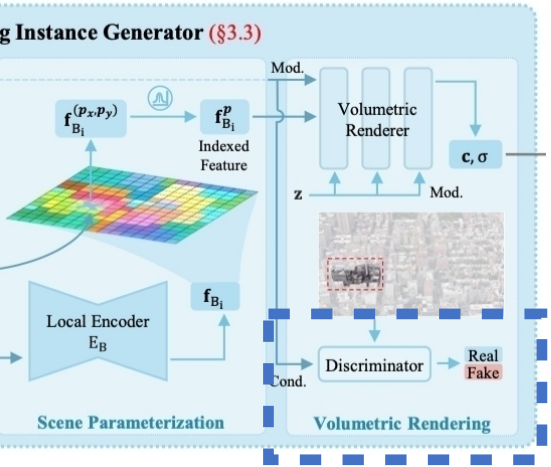
$\mathcal{O}(x)$: position code

$N_P^L - 1$:frequency range

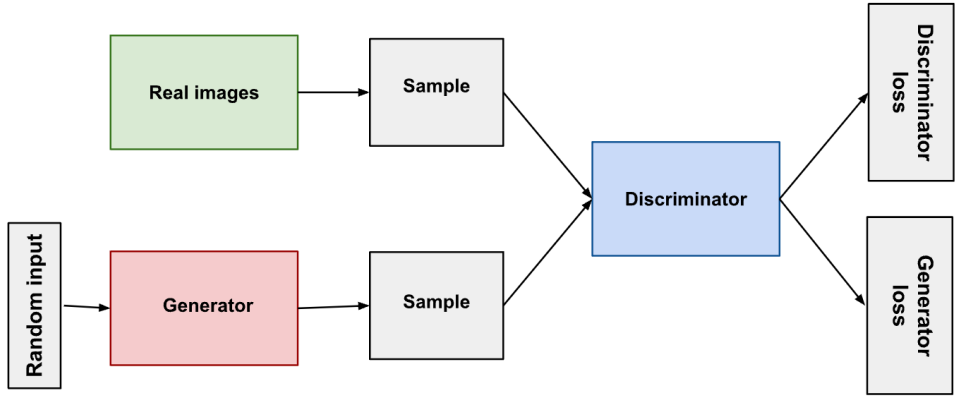
Position encode

City Background vs Building Generator - Diff GAN Loss





$$\ell_G = \lambda_{L1}\|\hat{\mathbf{I}}_G - \mathbf{I}_G\| + \lambda_P \mathcal{P}(\hat{\mathbf{I}}_G, \mathbf{I}_G) + \lambda_G \mathcal{G}(\hat{\mathbf{I}}_G, \mathbf{S}_G) \quad (6)$$



$$\ell_B = \mathcal{G}(\hat{\mathbf{I}}_{B_i}, \mathbf{S}_{B_i}) \quad (11)$$

ℓ_G: Total generation loss, λ_{L1}, λ_P, λ_G: Weight factor, $\hat{\mathbf{I}}_G$: Generated image, $\mathbf{I}_G$: Real image, $\mathbf{S}_G$: Semantic map, $\|\|$: L1 Loss, P: perceptual loss, $\mathcal{G}$: GAN loss

Compositor

$$\mathbf{I}_C = \hat{\mathbf{I}}_G \mathbf{M}_G + \sum_{i=1}^{n} \hat{\mathbf{I}}_{B_i} \mathbf{M}_{B_i} \qquad (12)$$

$\mathbf{I}_C$: Generated complete city image, $\hat{\mathbf{I}}_G$: Background image, $\mathbf{M}_G$:Background semantic mask

$\hat{\mathbf{I}}_{B_i}$: Building instance image, $\mathbf{M}_{B_i}$: Semantic mask of the building

Thoughts

unity

Q&A