PerspectiveNet: A Scene-consistent Image Generator for New View Synthesis in Real Indoor Environments

Authors: David Novotny, Benjamin Graham, Jeremy Reizenstein

Venue: Neural Information Processing Systems 33 (NeurIPS 2019)

Presenter: Surendar Devasundaram

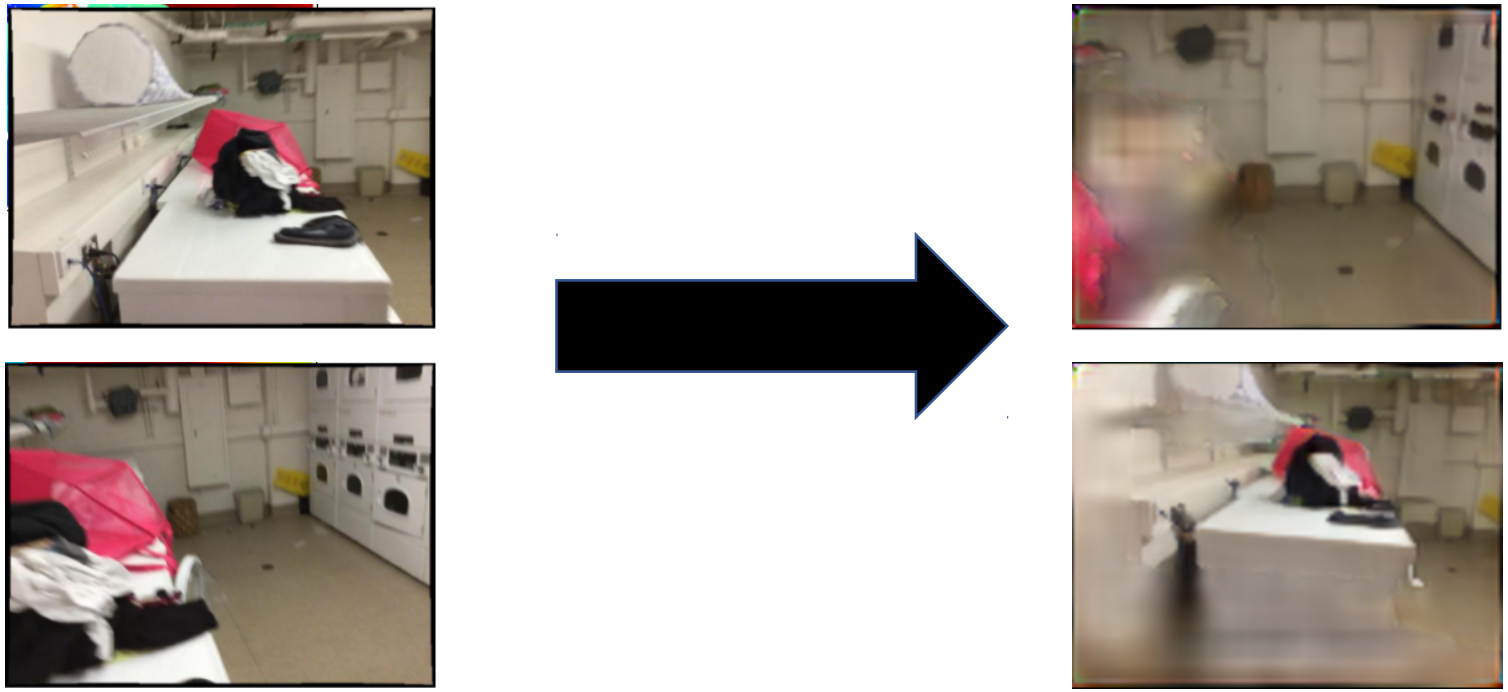
Faculty: Dr. Yogesh Singh Rawat

Outline

- Introduction
- Related Works
- Problem Statement
- Proposed Method
- Experiment
- Conclusion

Introduction

- New View Synthesis

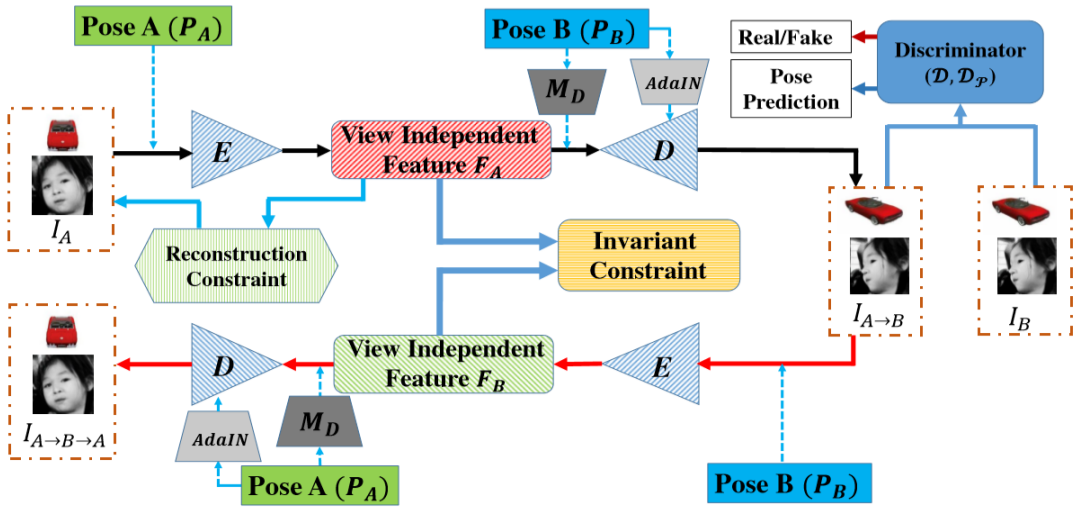


Reference View

New View

Related Works

- View Independent Generative Adversarial Network for Novel View Synthesis



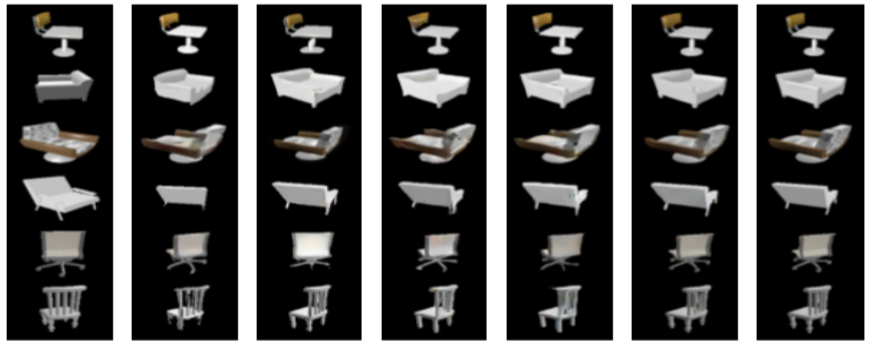


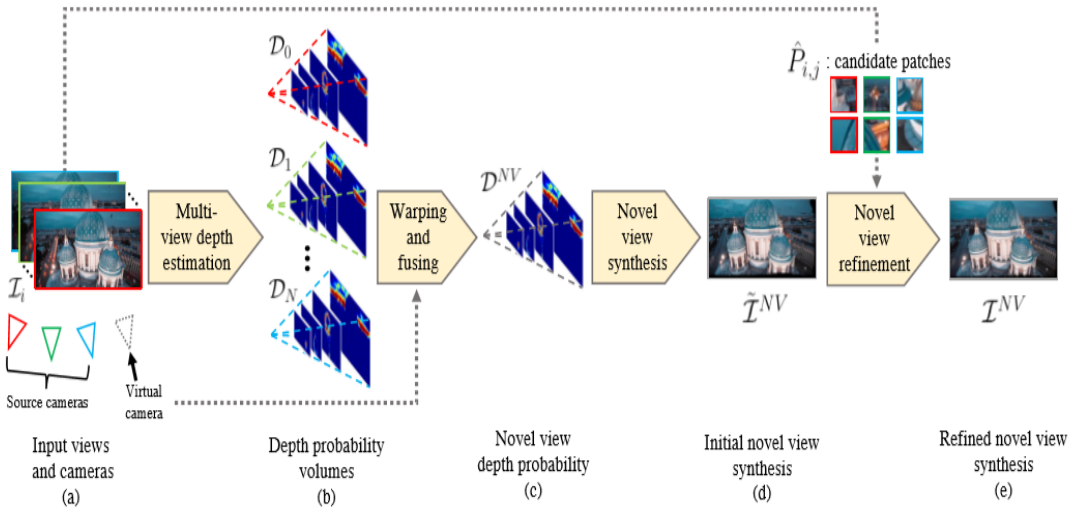
Figure 2. Results of ablation experiments. (a) is the input and (g) is the ground truth. (b)-(f) are synthesized by VI-GAN (w/o VI), VI-GAN (w/o Pose), VI-GAN (w/o Coord), VI-GAN (w/o AdaIN) and VI-GAN respectively. The images are with size 128×128.

Encoder $$F_A = E(I_A \oplus P_A)$$

Decoder $$I_{A\to B} = D(F_A \oplus M_D(P_B))$$

Related Works

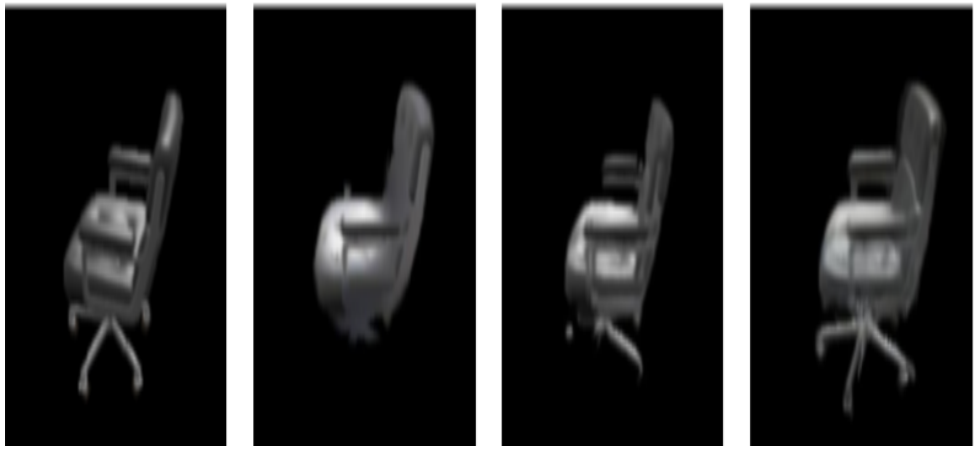
- Extreme View Synthesis





Problem Statement

- Synthesize new view for realistic indoor scene



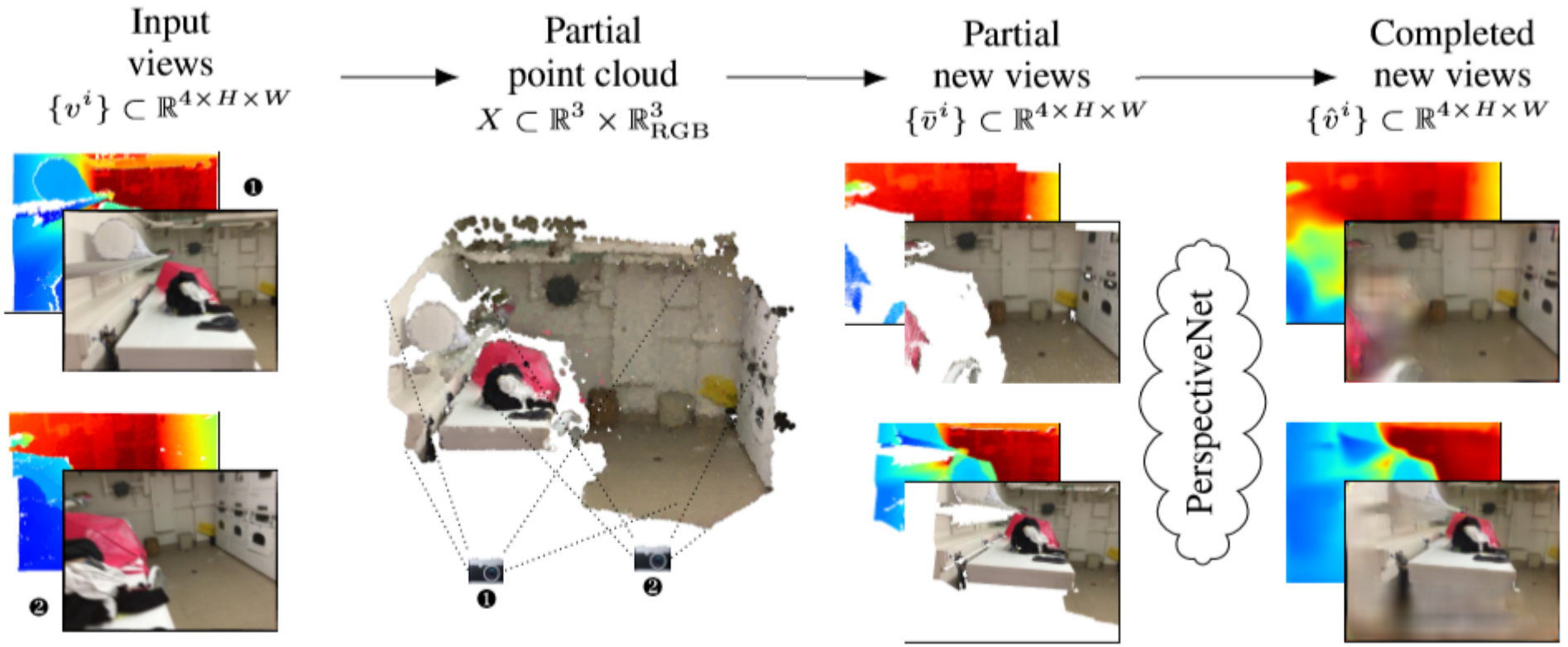
Category Specific images



Real Indoor scene

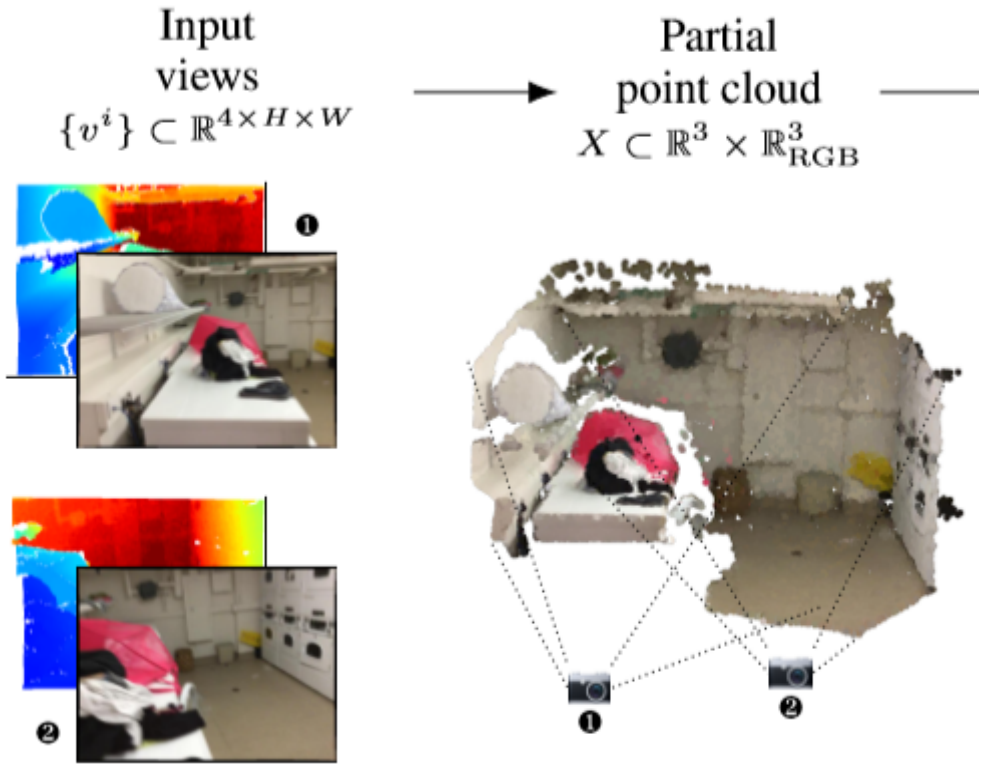
Proposed Method

- Overview



Proposed Method

- Generation of 3D Point Cloud
  - Input View $\{v^i\}_{i=1}^{N_{\text{ref}}}, v^i \in \mathbb{R}^{4\times H\times W}$
  - Input Camera Parameters
    - Extrinsic parameter $g^i \in SE(3)$
    - Intrinsic parameter $K^i \in \mathbb{R}^{4\times 4}$
    - Depth $d_u \in \mathbb{R}$
  - Pixel to Point Cloud Projection $x_u^i \sim (K^i g^i)^{-1}[u_1, u_2, d_u, 1]^T$



Proposed Method

- Generation of Partial Render Image
  - Project each point in X into camera frame

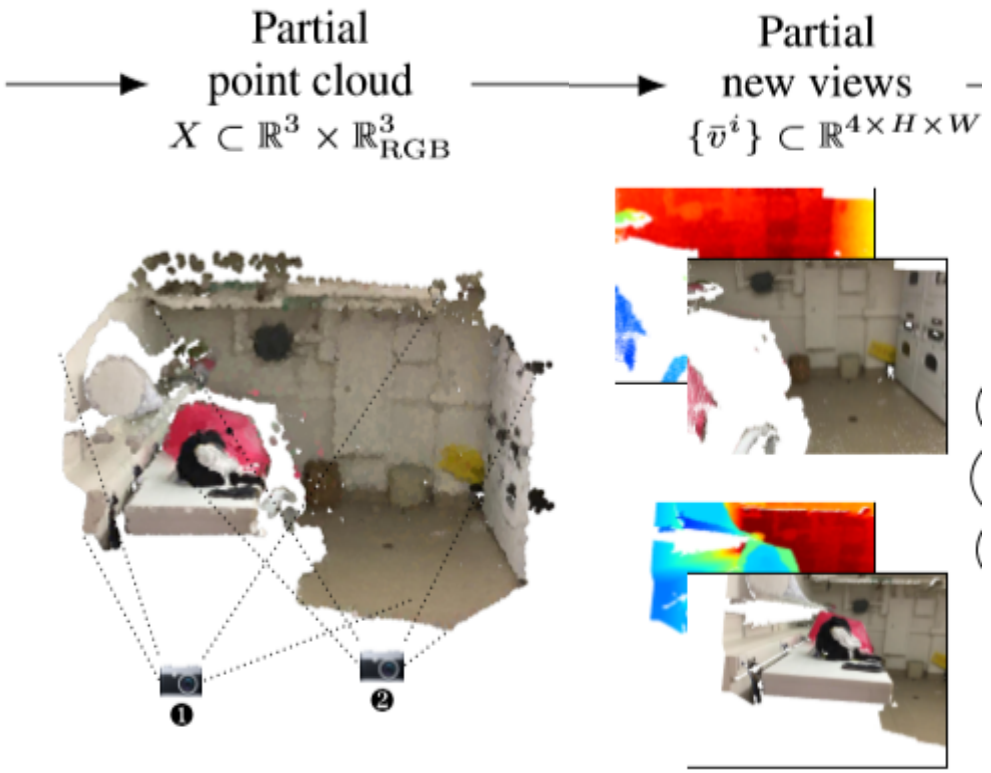
$$[u_{1j} \;\; u_{2j} \;\; d_j \;\; 1]^T \sim Kg[x_j^T \;\; 1]^T$$

  - Calculate weight for floating point coordinates

$$w_{a,b,j} = (1 - |u_{1j} - a|)(1 - |u_{2j} - b|)\exp(-d_j)$$

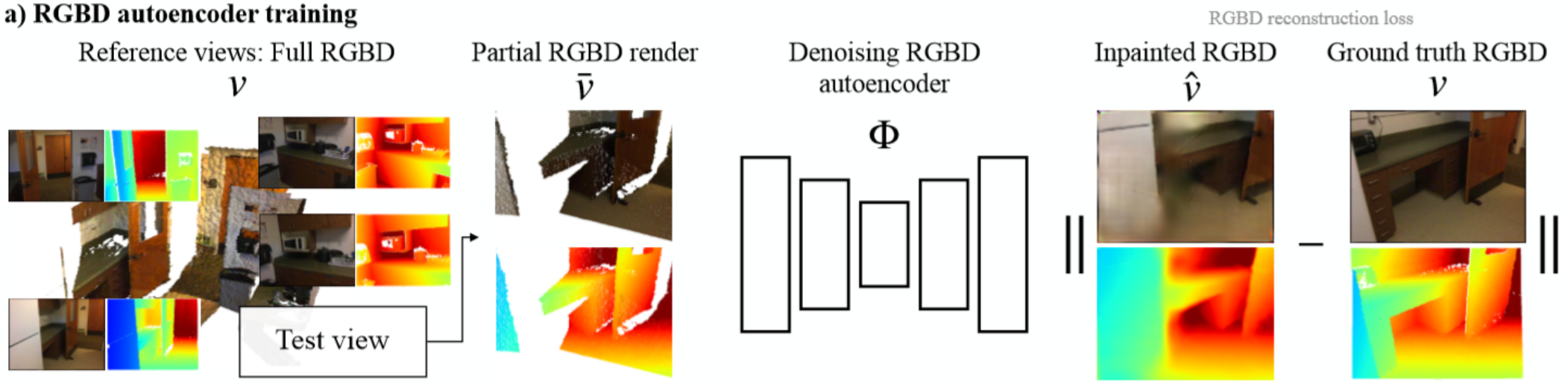
$$\text{if } |u_{1j} - a| < 1 \text{ and } |u_{2j} - b| < 1$$

  - Color each pixel in the image



Proposed Method

- Inpainting the partial render image
  - Using denoising RGBD autoencoder



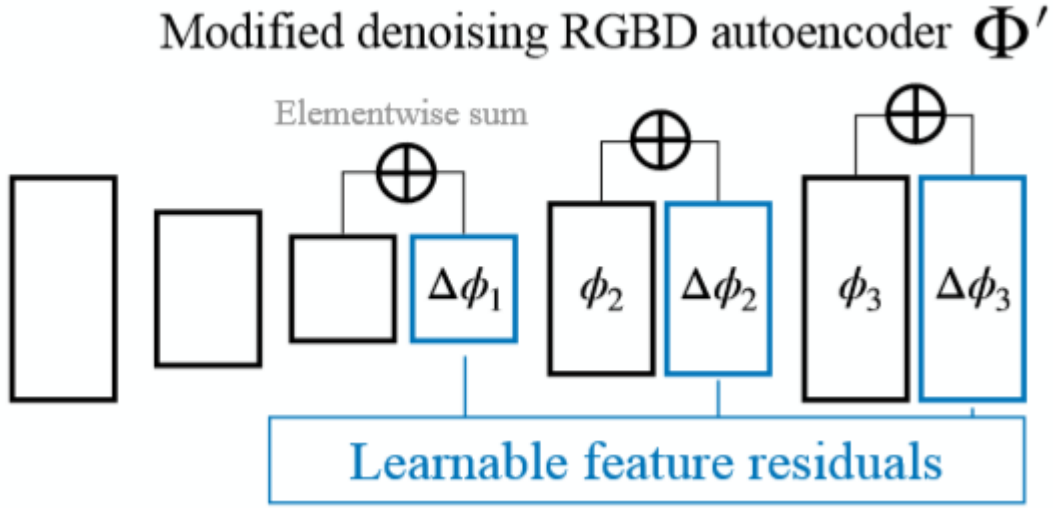
  - Training Dataset: Partial Renders from ScanNet
  - 150 epochs using Adam optimizer
  - Learning rate: 10^-5 decaying 10-fold after visiting 100k minibatches

Proposed Method

- Inpainting the partial render image
  - Drawbacks of Using denoising RGBD autoencoder
    - Cannot handle large missing areas
    - Inpainting inconsistent across different views

Proposed Method

- How Scene-consistent Inpainting is achieved?
  - Latent image parameterization



$$\Phi'(\bar{v}, \Delta\phi(\bar{v})) = \Phi_L(\ldots \Phi_1(\Phi_0(\bar{v}) + \Delta\phi_1) \ldots + \Delta\phi_L)$$

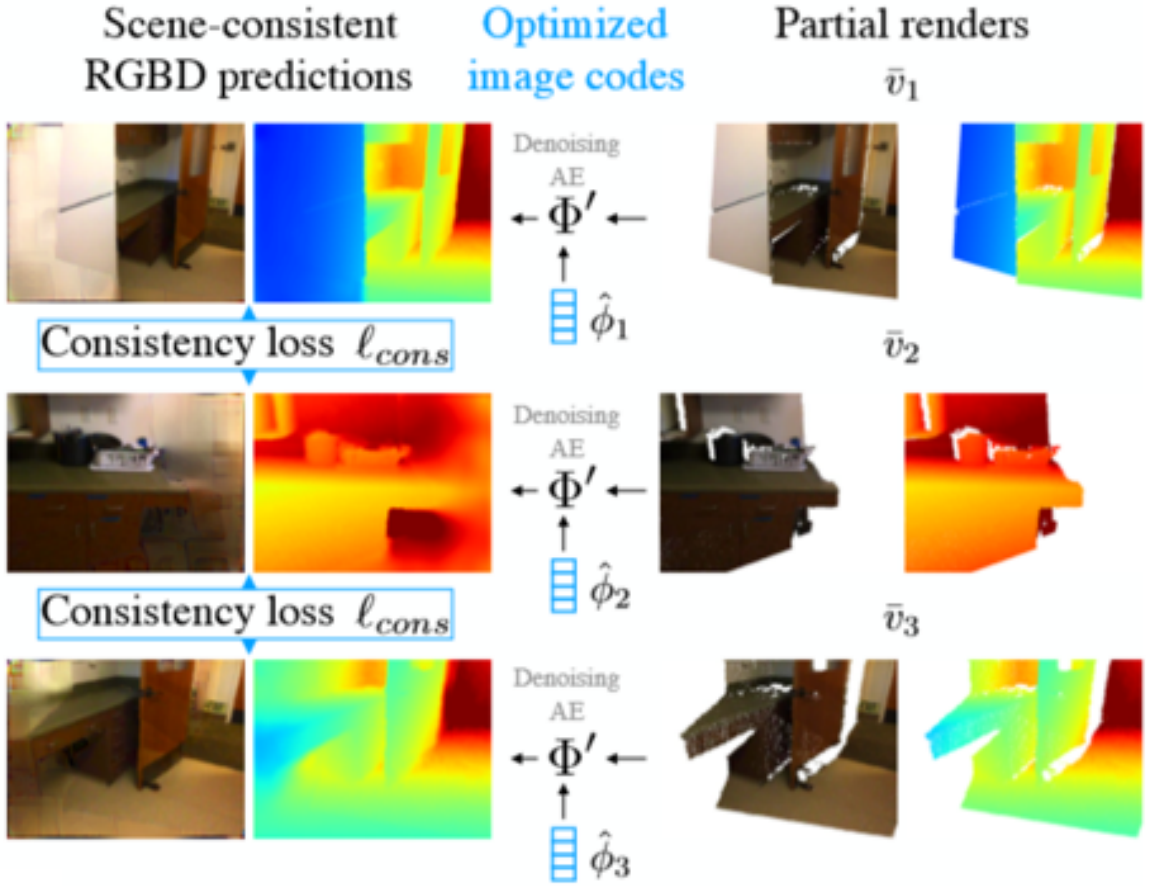
Proposed Method

- How Scene-consistent Inpainting is achieved?
  - Choice of consistency loss
    - Reprojection Consistency Loss

$$\ell_{\text{cons}}\left(\hat{v}^i \,\middle|\, V \backslash \{\hat{v}^i\}\right) = \sum_{u \in \Omega(\hat{v}^i) \cap \Omega(\check{v}^i)} h(\hat{v}^i_u, \check{v}^i_u),$$

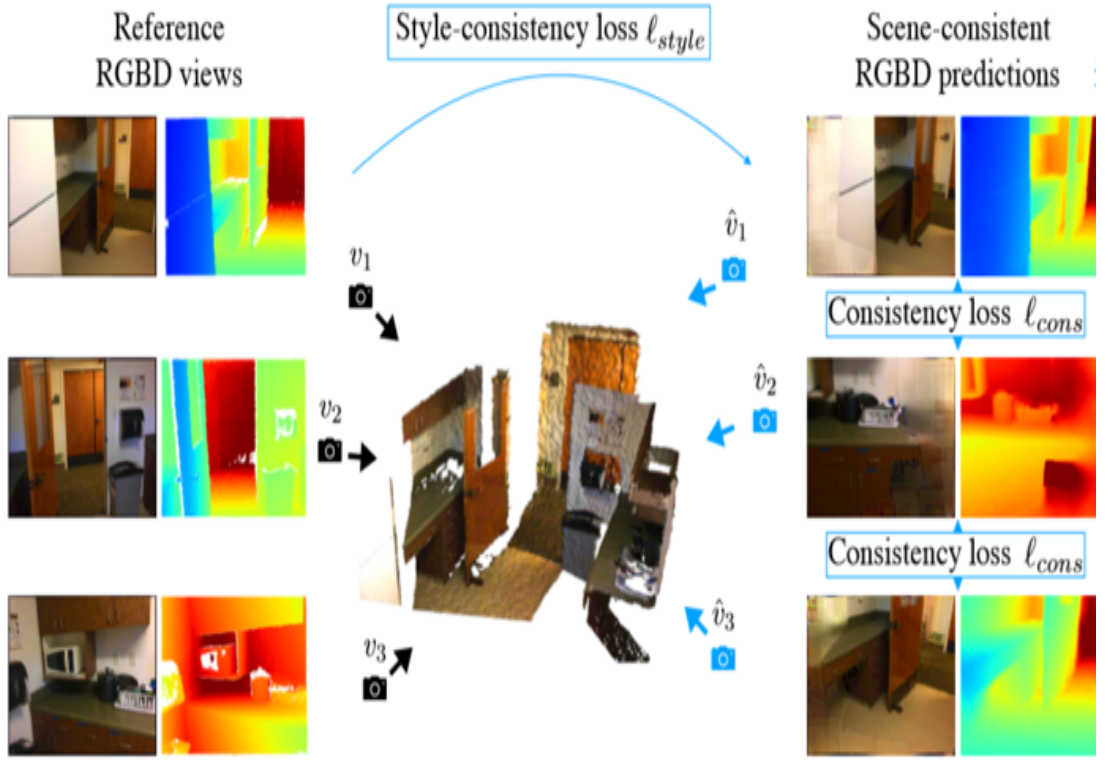
$$h(a,b) = \delta \sum_{\mathbf{c}=1}^{6} \left(\sqrt{1 + \delta^{-1}(a_{\mathbf{c}} - b_{\mathbf{c}})^2} - 1 \right)$$

$$\ell_R = \sum_{i=1}^{N_{\text{test}}} \left[\sum_{u \in \Omega(\bar{v}^i_u)} h(\bar{v}^i_u, \hat{v}^i_u) + \sum_{u \in \Omega(\hat{v}^i_u)} h(\hat{v}^i_{u,t=0}, \hat{v}^i_u) \right]$$



Proposed Method

- How Scene-consistent Inpainting is achieved?
  - Choice of consistency loss
    - Style Consistency loss



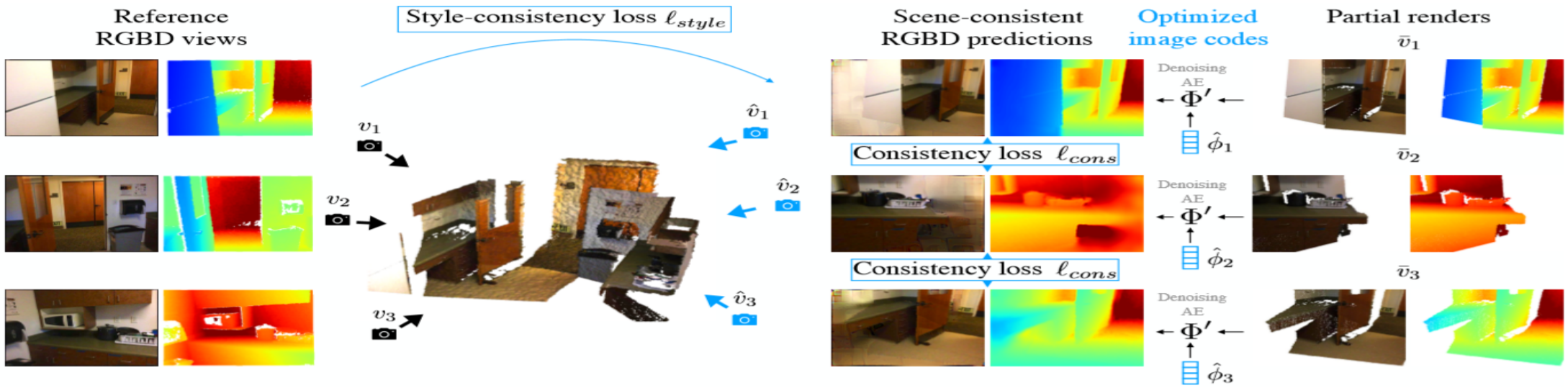
$$\ell_{\text{style}}\left(\{\hat{v}^i\}_{i=1}^{N_{\text{test}}} \,\middle|\, \{v^i\}_{i=1}^{N_{\text{ref}}}\right) = \sum_{l\in 1,2,3} \left| \sum_{i=1}^{N_{\text{ref}}} \frac{\Psi_l(v^i)\Psi_l(v^i)^T}{N_{\text{ref}} H_l W_l} - \sum_{i=1}^{N_{\text{test}}} \frac{\Psi_l(\hat{v}^i)\Psi_l(\hat{v}^i)^T}{N_{\text{test}} H_l W_l} \right|_1$$

$$\ell_{\text{TV}} = \frac{1}{N_{\text{test}} W H} \sum_{i=1}^{N_{\text{test}}} \sum_{u_1,u_2} |c^i_{(u_1,u_2+1)} - c^i_{(u_1,u_2)}| + |c^i_{(u_1+1,u_2)} - c^i_{(u_1,u_2)}|$$

Proposed Method

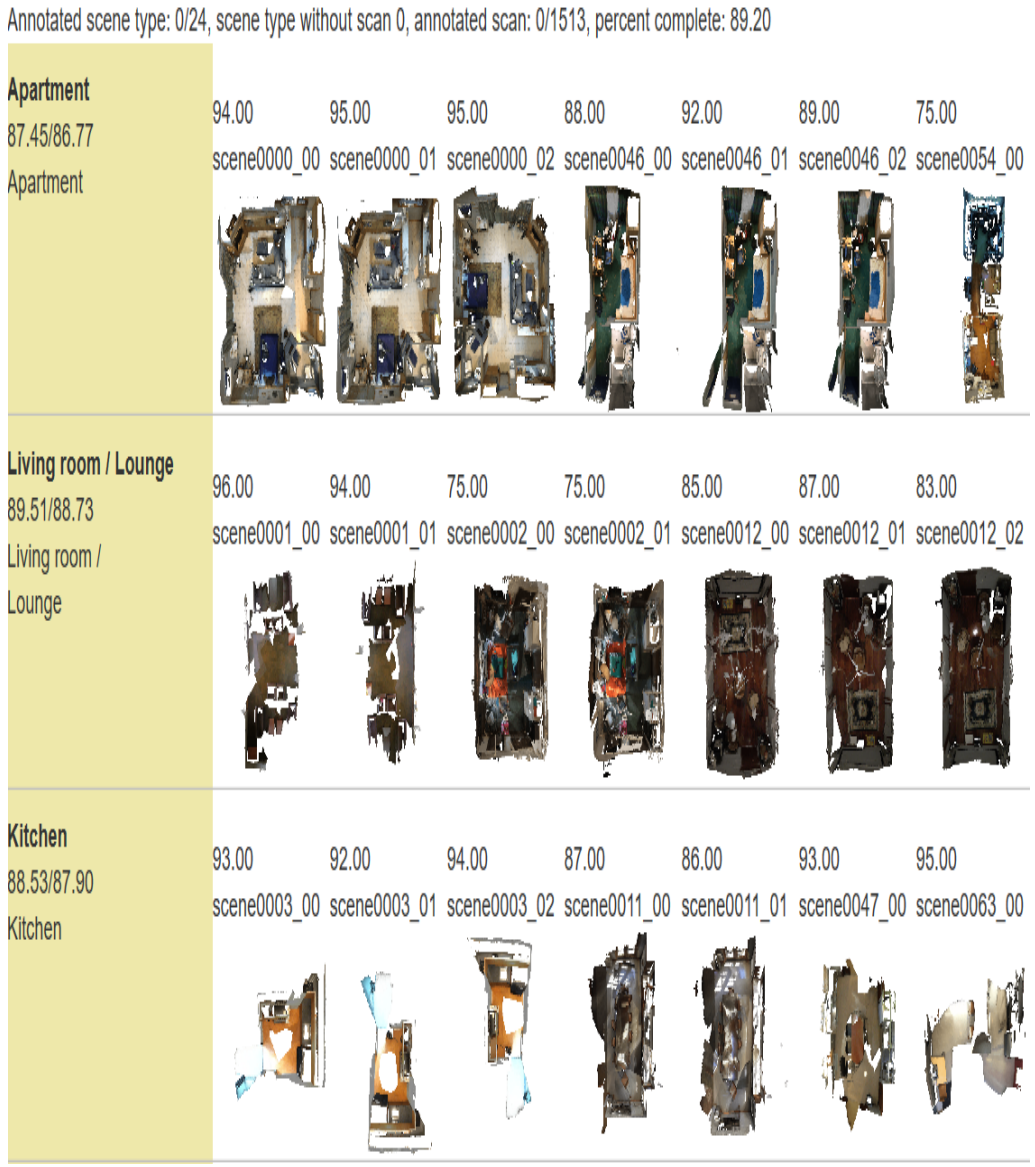
- How Scene-consistent Inpainting is achieved?
  - Scene level optimization
    - Choice of consistency loss

$$\mathcal{L} = \underset{\{\Delta\phi(\bar{v}^i)\}_{i=1}^{N_{\text{test}}}}{\arg\min}\ \ell_{\text{style}}\left(\{\hat{v}^i\}_{i=1}^{N_{\text{test}}} \,\middle|\, \{v^i\}_{i=1}^{N_{\text{ref}}}\right) + \sum_{i=1}^{N_{\text{test}}} \ell_{\text{cons}}\left(\hat{v}^i \,\middle|\, V \backslash \{\hat{v}^i\}\right) \qquad \hat{v}^i = \Phi'(\bar{v}^i, \Delta\phi(\bar{v}^i))$$



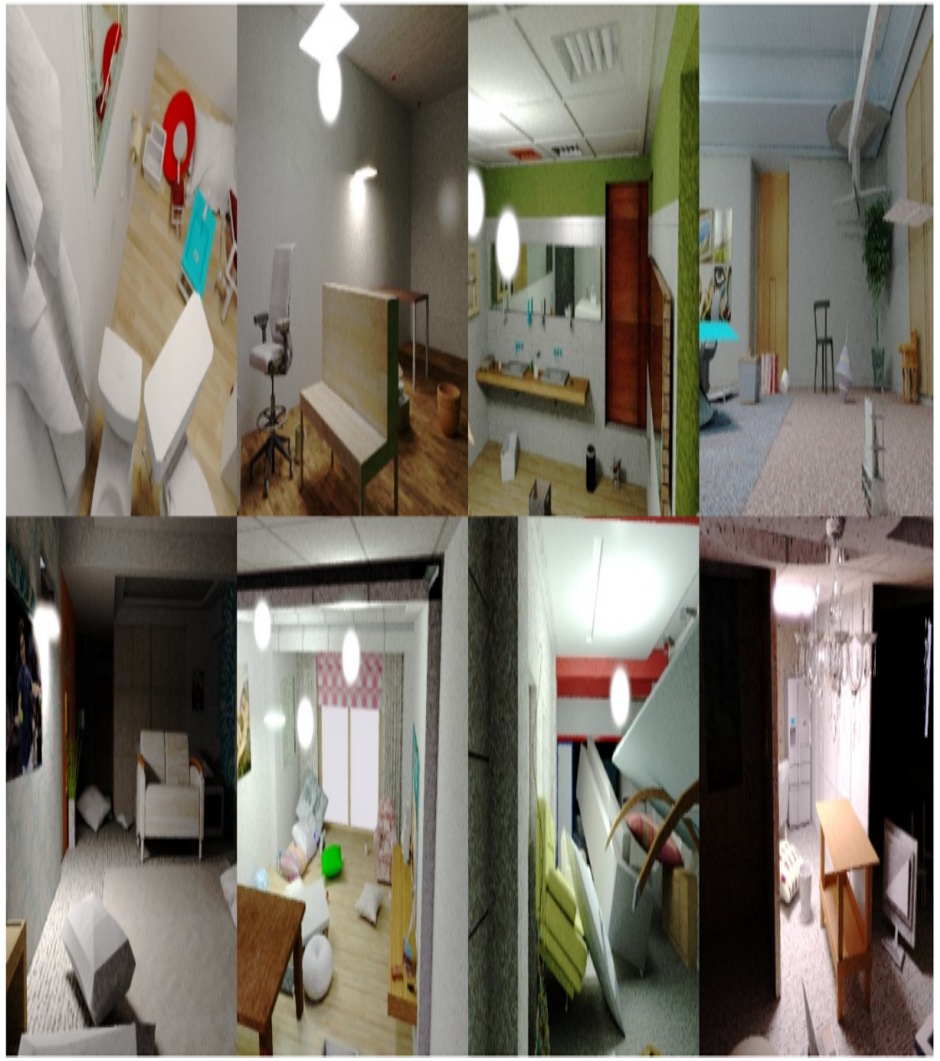
Experiment

- Dataset
  - ScanNet
    - 3D dataset of real indoor scenes
    - Training scenes: 1500
    - Test scenes: 100
    - Annotation: Camera extrinsic and intrinsic



Experiment

- Dataset
  - SceneNet
    - Synthetic dataset
    - Training scenes: 33000
    - Test scenes: 1000
    - Annotation: Camera extrinsic and intrinsic



Experiment

- Evaluation Protocol
  - For a given test scene
    - 4 reference view with knowledge of their RGBD and camera parameters.
    - 8 reference view with only camera parameters.
    - For a test view
      - Color and depth channels are generated and compared with the corresponding ground truth frames.
    - ScanNet – All 100 test scenes, SceneNet – Randomly selected 100 test scenes

Experiment

- Evaluation Metrics
  - Per-pixel L1 error
  - Peak-signal-to-noise-ratio
  - LPIPS

(a) **ScanNet**

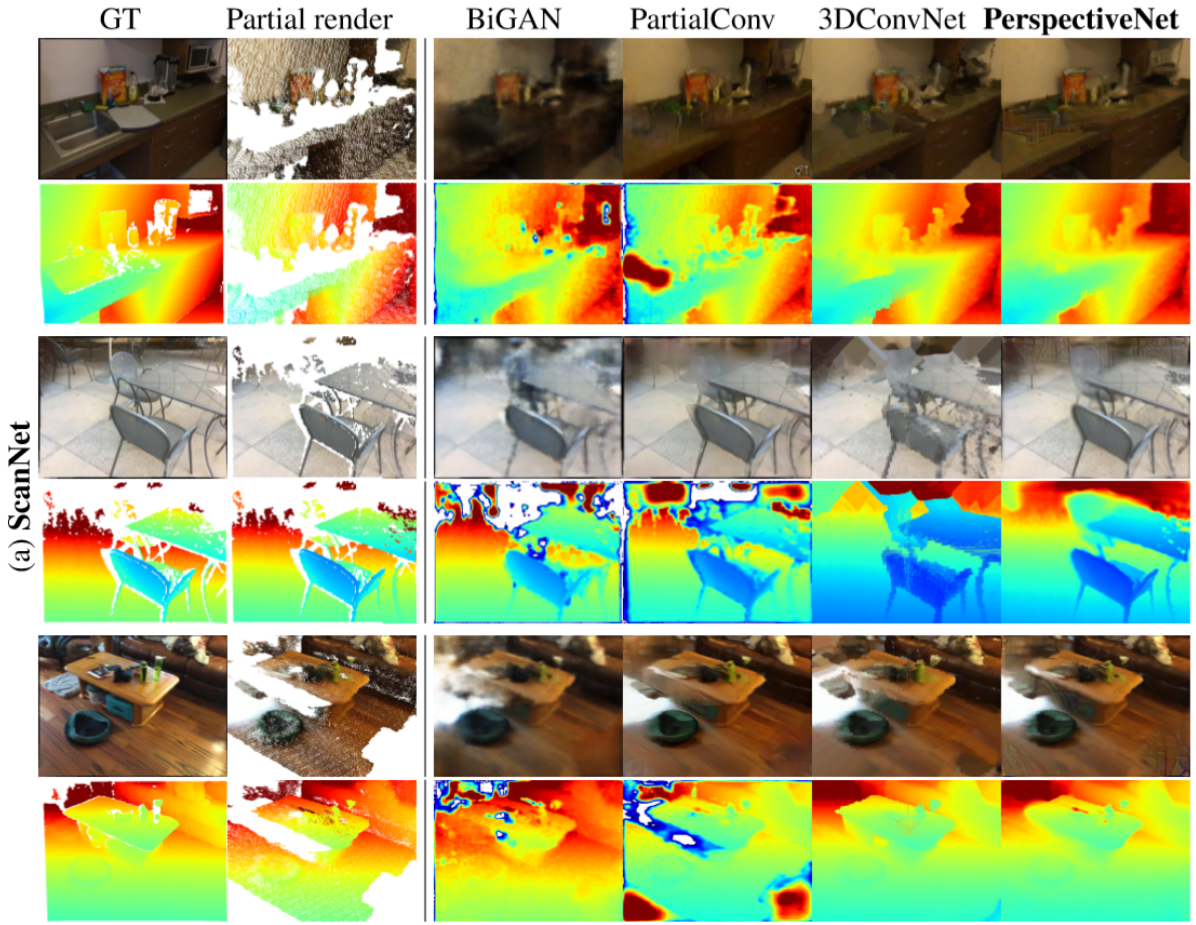
Method	Color metrics			Depth metrics			
	$\ell_1^{RGB} \downarrow$	PSNR ↑	LPIPS ↓	$\ell_1^{D}[m] \downarrow$	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
PerspectiveNet	68.022	13.762	**0.422**	**0.115**	**0.352**	**0.411**	**0.471**
PerspectiveNet w/o opt	**66.511**	**13.986**	0.426	0.120	0.188	0.230	0.279
PartialConv [26]	93.604	11.374	0.461	0.750	0.194	0.236	0.283
3DConvNet	78.590	12.190	0.531	0.138	0.301	0.359	0.426
BiGAN [45]	77.313	12.742	0.523	0.215	0.169	0.212	0.265

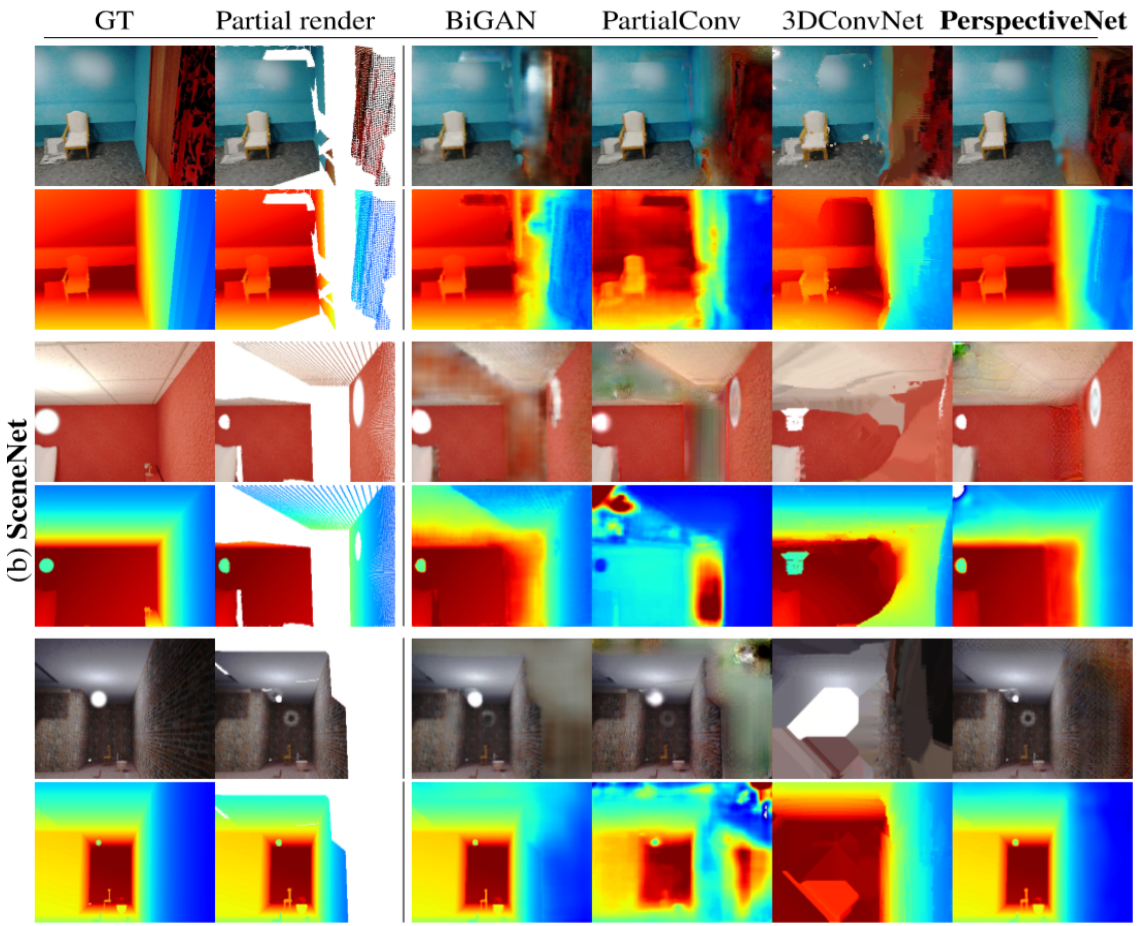
(b) **SceneNet**

Method	$\ell_1^{RGB} \downarrow$	PSNR ↑	LPIPS ↓	$\ell_1^{D}[m] \downarrow$	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
PerspectiveNet	49.698	15.687	**0.424**	**0.219**	**0.366**	**0.431**	**0.494**
PerspectiveNet w/o opt	**48.521**	**16.324**	0.442	0.227	0.101	0.125	0.155
PartialConv [26]	76.470	12.377	0.481	1.846	0.008	0.010	0.013
3DConvNet	75.942	12.614	0.570	0.653	0.040	0.050	0.062
BiGAN [45]	55.815	15.112	0.485	0.249	0.319	0.375	0.431

Experiment

- Discussion of results





Conclusion

- Novelty is use of denoising autoencoder and finding latent representations that are consistent for different views
- Addresses the problem on realistic, large scale indoor scene dataset.

THANK YOU