Novel View Synthesis from only a 6-DoF Camera Pose by Two-stage Networks

Xiang Guo, Bo Li, Yuchao Dai *, Tongxin Zhang, Hui Deng

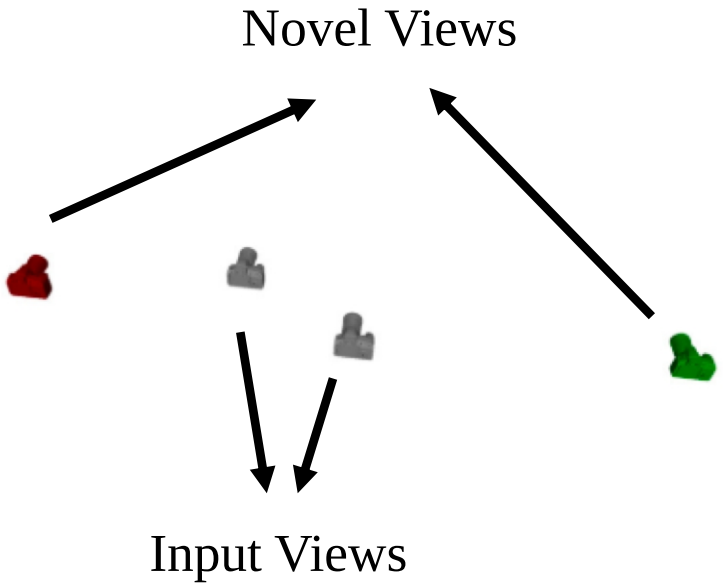
School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China

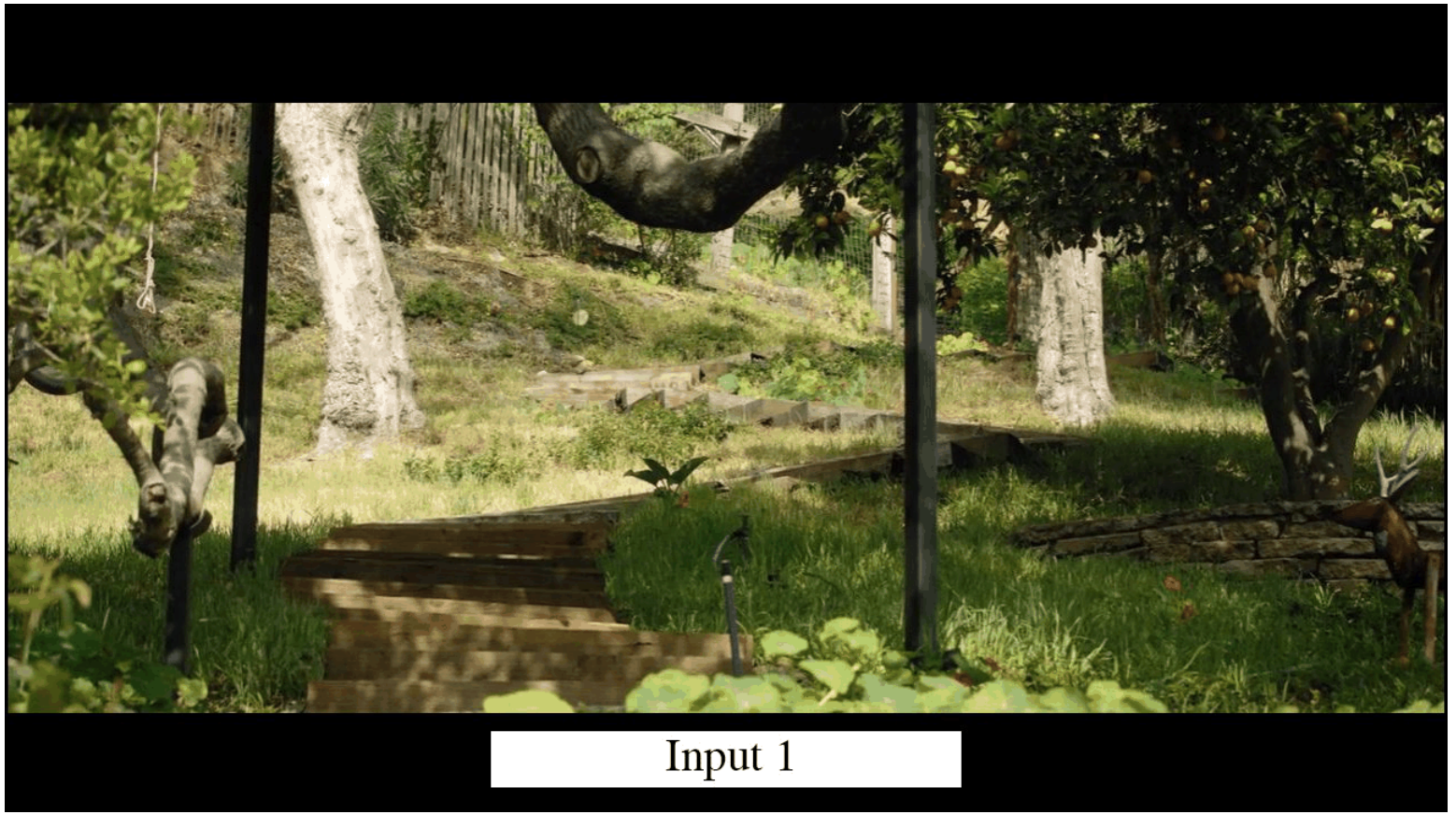
Content

- Introduction
- Approach
- Results
- Ablation Studies
- Conclusion

Introduction

The task of Novel View Synthesis (NVS)





Choi et al., ICCV'19

Introduction

The Dependency of NVS in Synthesizing Process

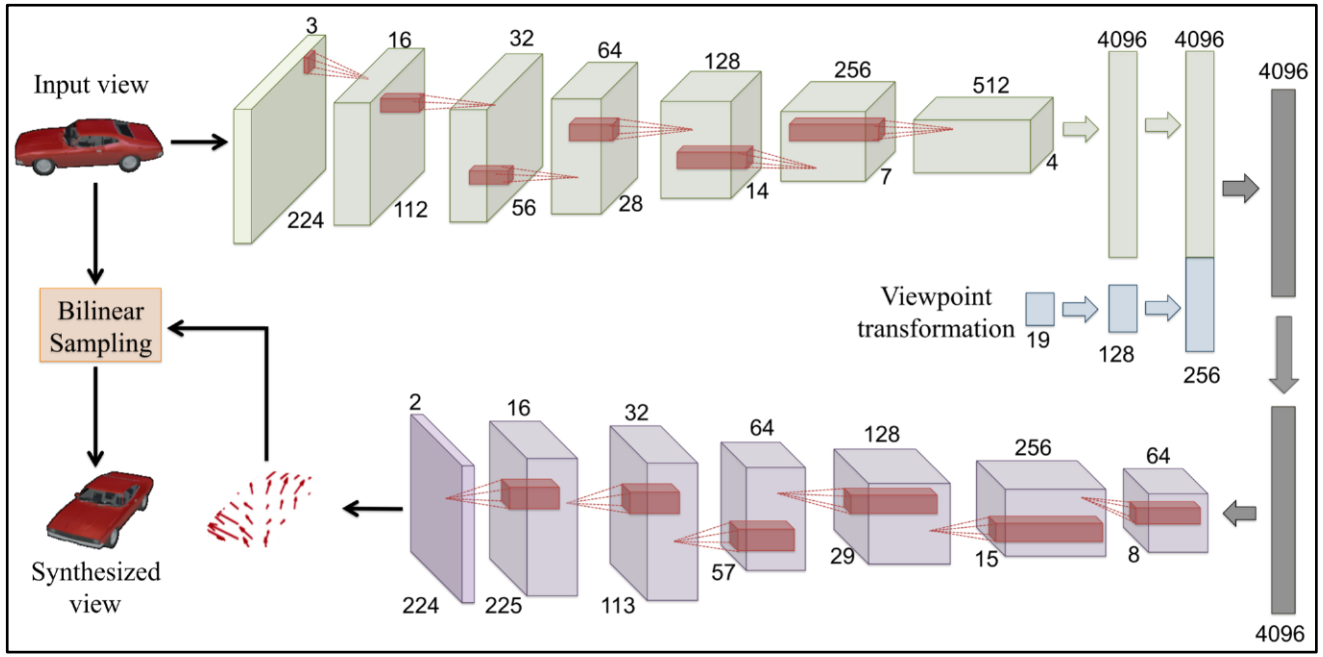
- Camera Pose

Introduction

The Dependency of NVS in Synthesizing Process

- Camera Pose

+ Input Images for Reference



Zhou et al., ECCV'16

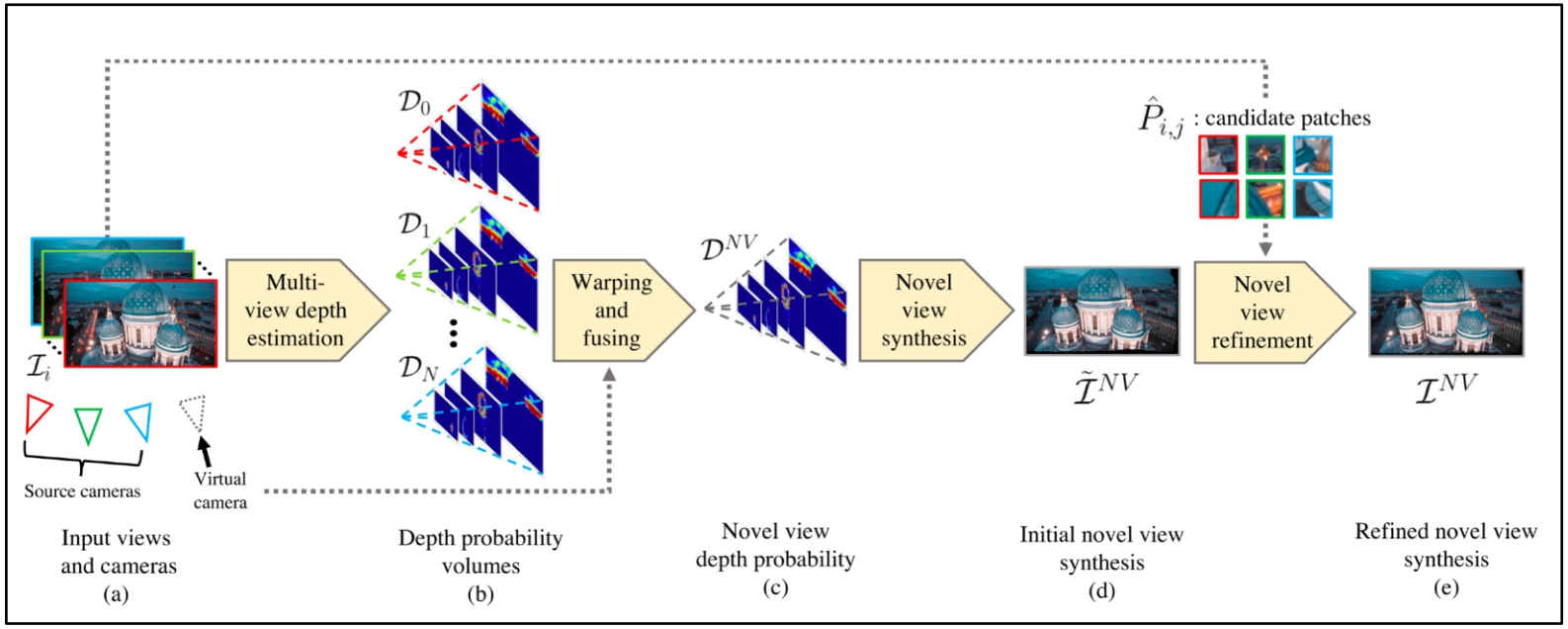
Introduction

The Dependency of NVS in Synthesizing Process

- Camera Pose

+ Input Images for Reference

+ Depth



Choi et al., ICCV'19

Introduction

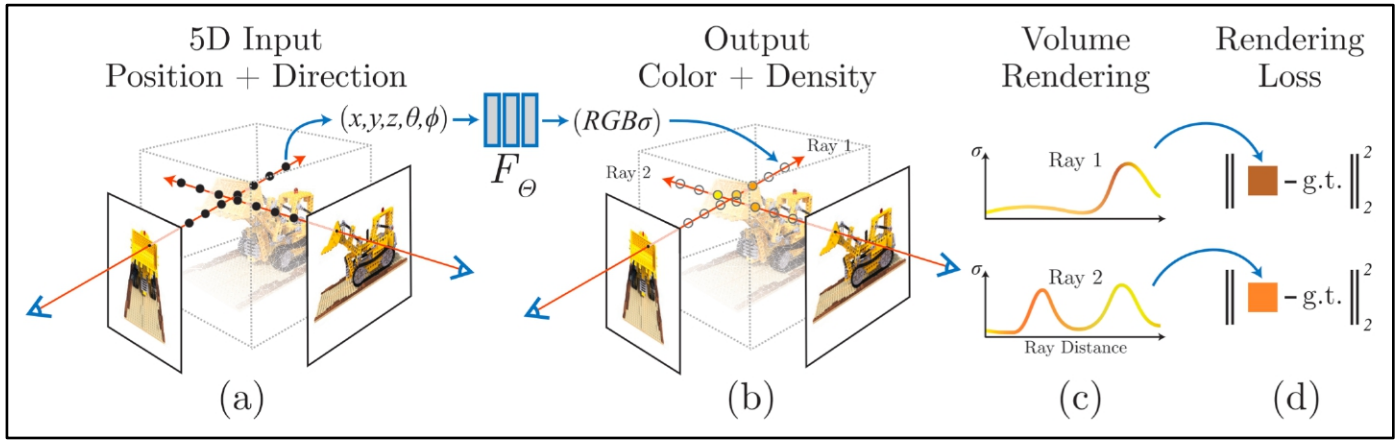
The Dependency of NVS

- Camera Pose

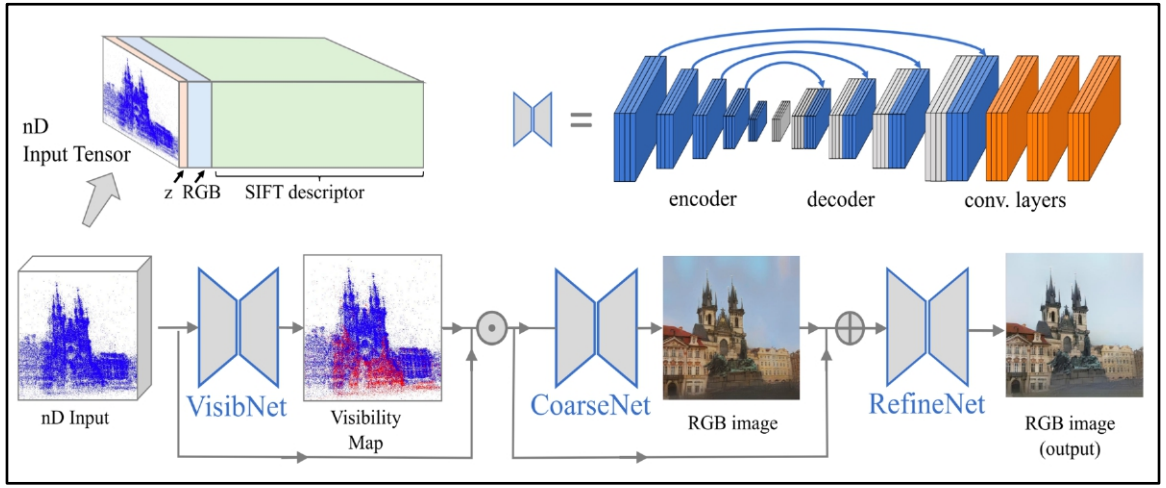
+ Input Images for Reference

+ Depth

+ 3D model



Mildenhall et al., ECCV'20



Pittaluga et al., CVPR'19

Riegler and Koltun, ECCV'20

Introduction

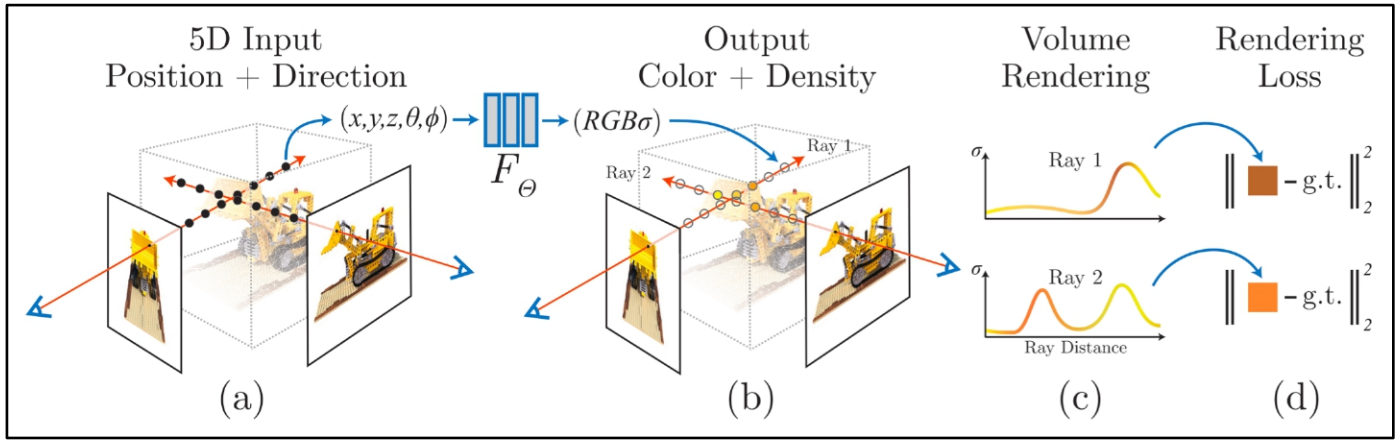
The Dependency of NVS

- Camera Pose

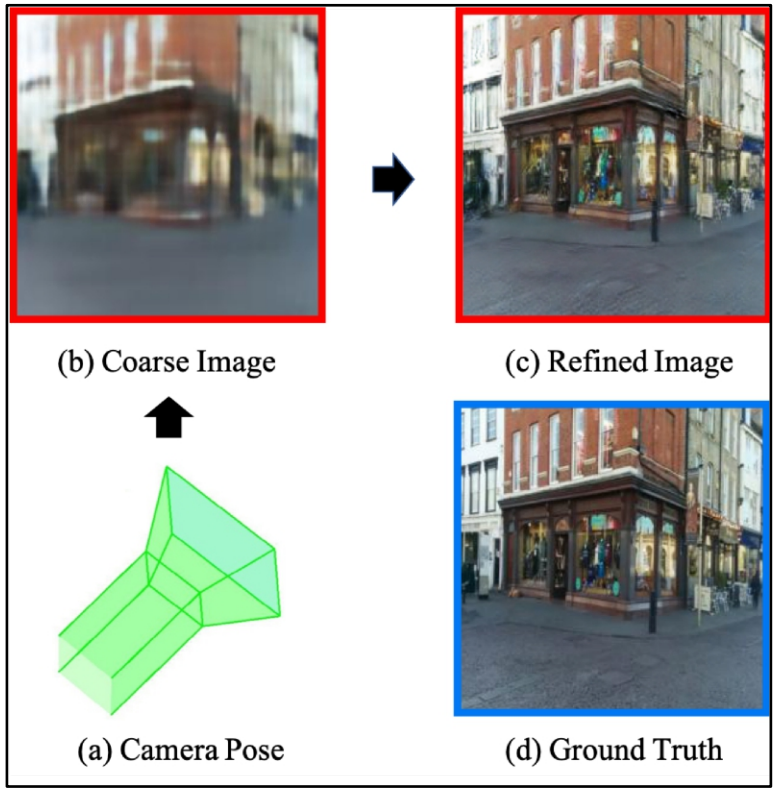
+ Input Images for Reference

+ Depth

+ 3D model



Mildenhall et al., ECCV'20



Only Pose as Input (Ours)

Treat rendering and scene representation as one task

Introduction

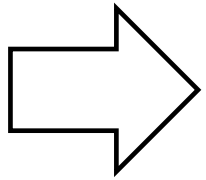
- Camera Pose

+ Input Images for Reference

+ Depth

+ 3D model

Limitation



- Accuracy
- Availability
- Memory Consumption
- Computation complexity

Introduction

- Camera Pose

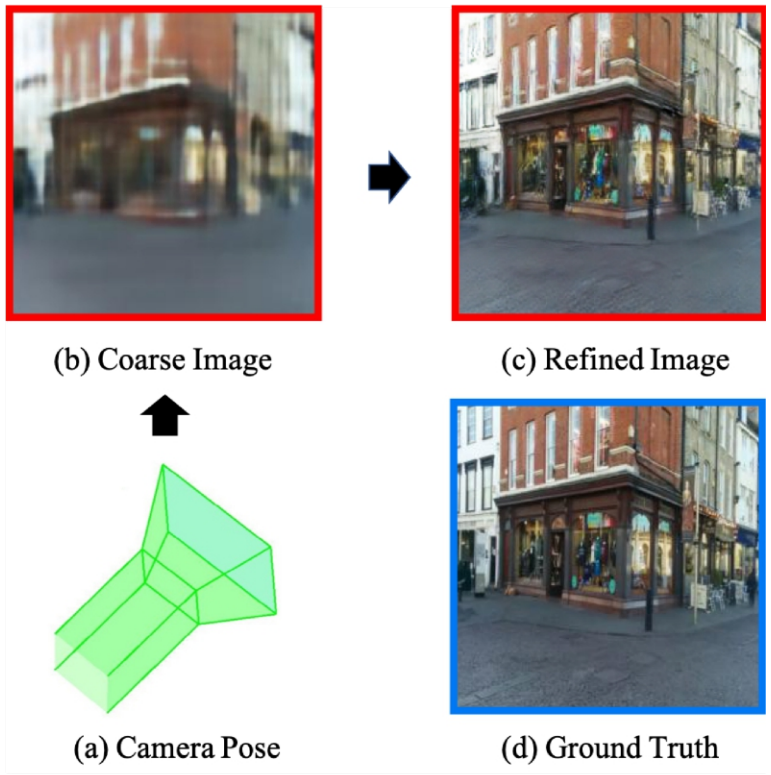
+ Input Images for Reference

+ Depth

+ 3D model

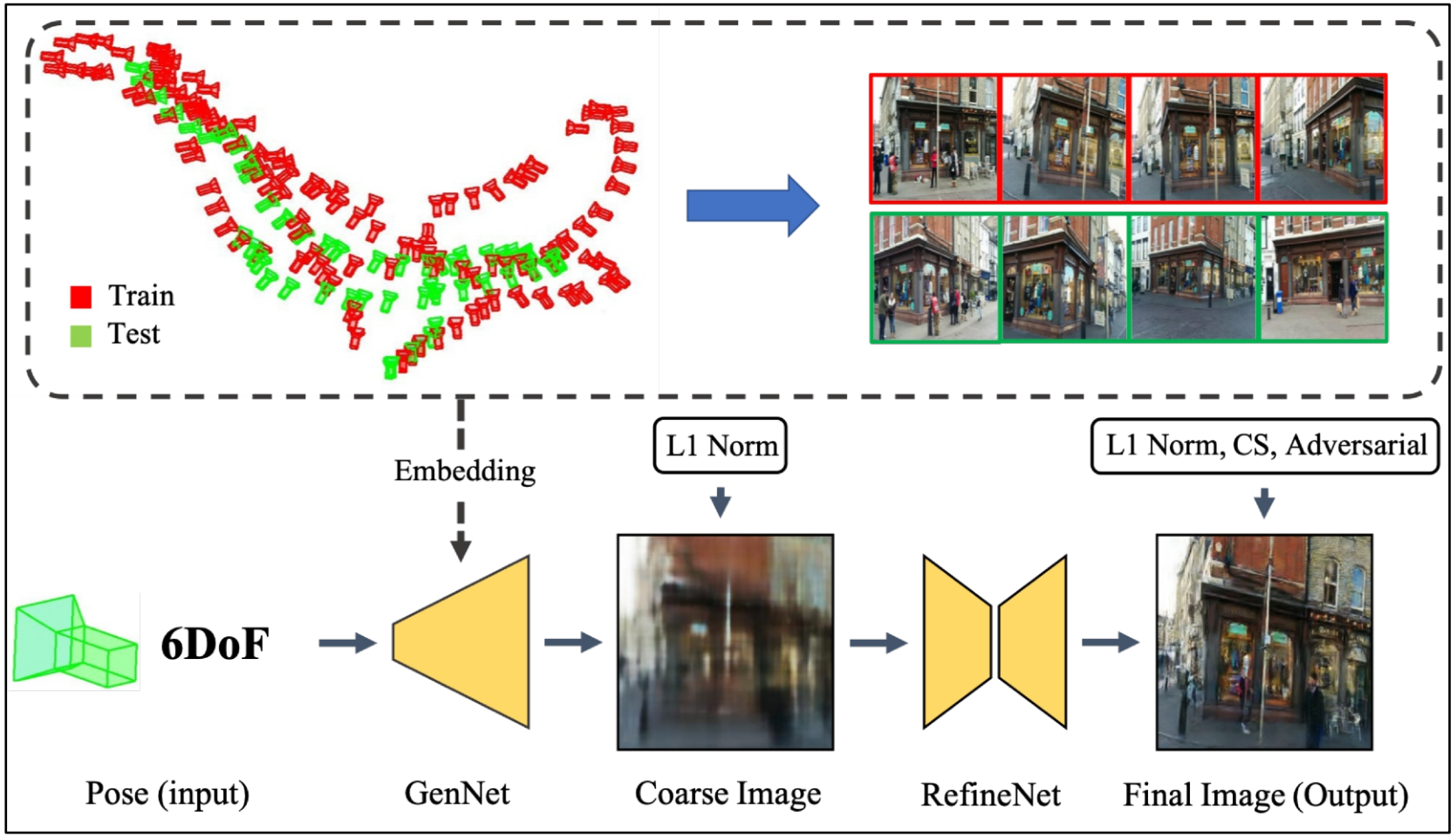
VS

- Camera Pose Only!



Only Pose as Input (Ours)

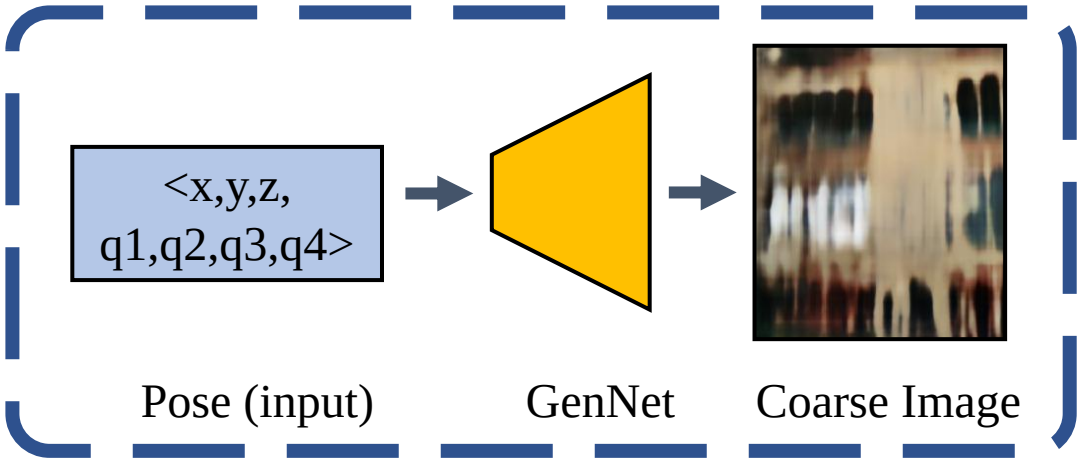
Approach - Overview

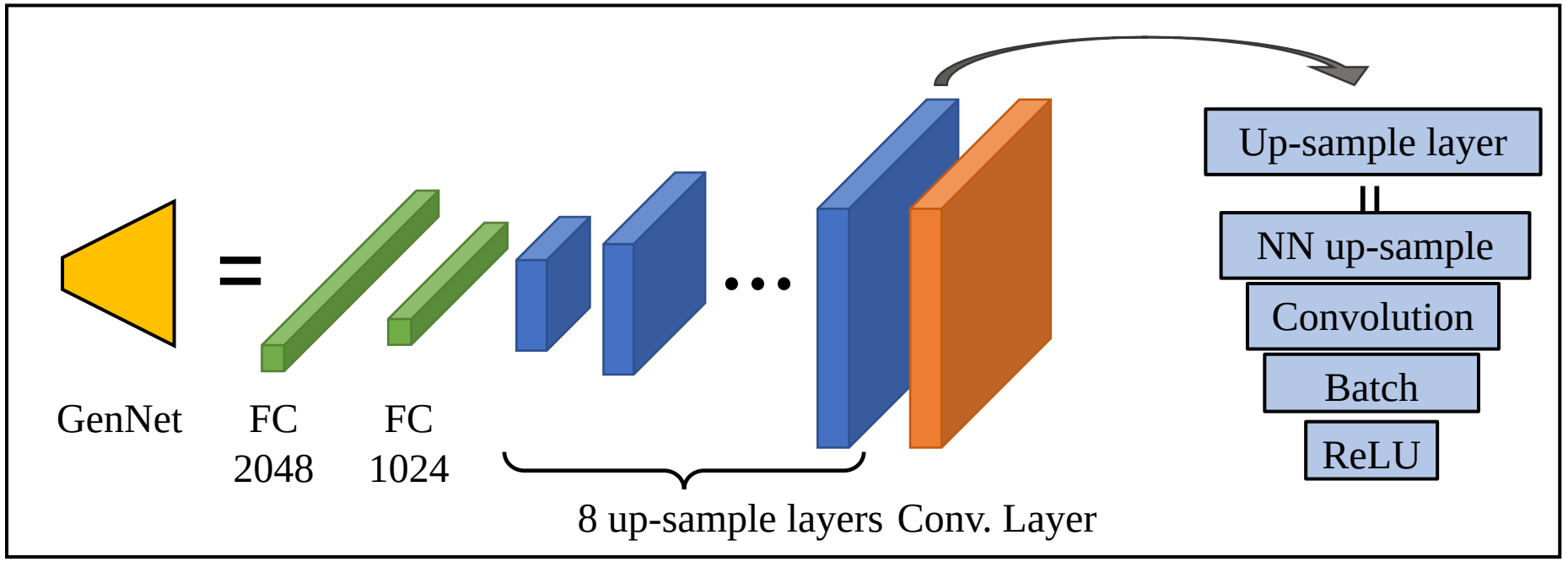


Overview of our method. The scene information is embedded by the network weights during Training

Approach

Stage-1: GenNet

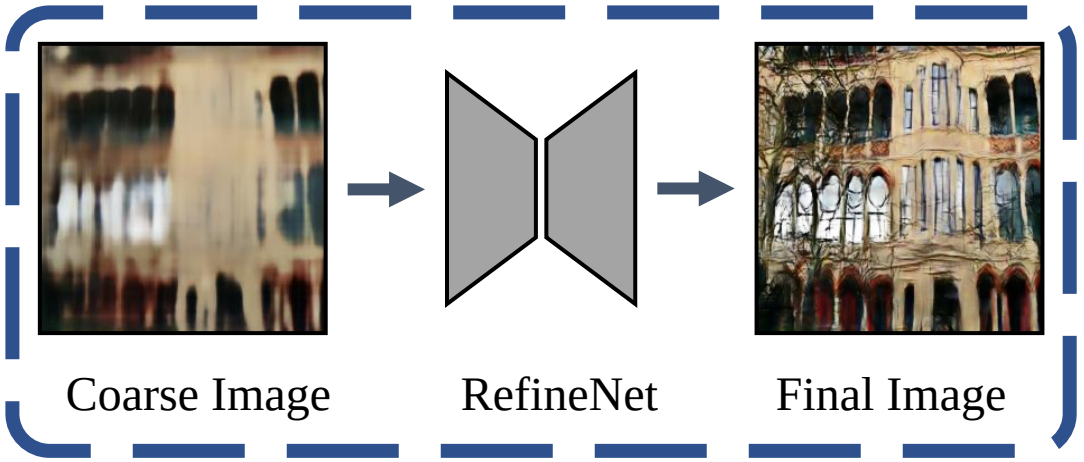




GenNet network structure. Expand pose in channel dimension

Approach

Stage-2: RefineNet



- U-Net Style, follow pix2pix [1]
- Loss: L1 norm, Perceptual Loss, Adversarial Loss

$$G^* = \arg\min_G \max_D \; l_{\text{cGAN}}(G, D) \; + \\ \lambda_1 l_{\text{L1}}(G) + \lambda_2 l_{\text{style}}(G) + \lambda_3 l_{\text{content}}(G)$$

[1] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 5967–5976.

Results

Examples on Cambridge Landmarks

OldHospital

GreatCourt

ShopFacade

KingsCollege

StMarysChurch

Street

Results

Results on Cambridge Landmarks [1]

	GreatCourt			KingsCollege			OldHospital		
	Coarse	Refined (PL)	Refined (w/o PL)	Coarse	Refined (PL)	Refined (w/o PL)	Coarse	Refined (PL)	Refined (w/o PL)
SSIM	0.4361	0.3916	0.3903	0.2953	0.2397	0.2370	0.1897	0.1300	0.1308
PSNR	11.5266	11.3886	11.3648	12.9879	12.5296	12.5288	12.4578	11.5834	11.5353
L1	0.2023	0.2050	0.2055	0.1706	0.1803	0.1800	0.1785	0.1977	0.1987
Brenner	191.3672	466.3329	471.1897	232.5756	744.3240	752.2070	180.5834	1360.8733	1368.9296
	ShopFacade			StMarysChurch			Street		
	Coarse	Refined (PL)	Refined (w/o PL)	Coarse	Refined (PL)	Refined (w/o PL)	Coarse	Refined (PL)	Refined (w/o PL)
SSIM	0.2495	0.1768	0.1372	0.2962	0.2172	0.2158	0.2255	0.1303	0.1372
PSNR	12.7433	11.9713	11.9403	12.8790	12.4334	12.3585	10.3989	9.7498	9.6893
L_1	0.1818	0.1951	0.1955	0.1777	0.1855	0.1867	0.2456	0.2633	0.2653
Brenner	113.3604	692.7051	697.9377	117.7310	538.1169	619.7828	109.2868	870.7560	997.1627

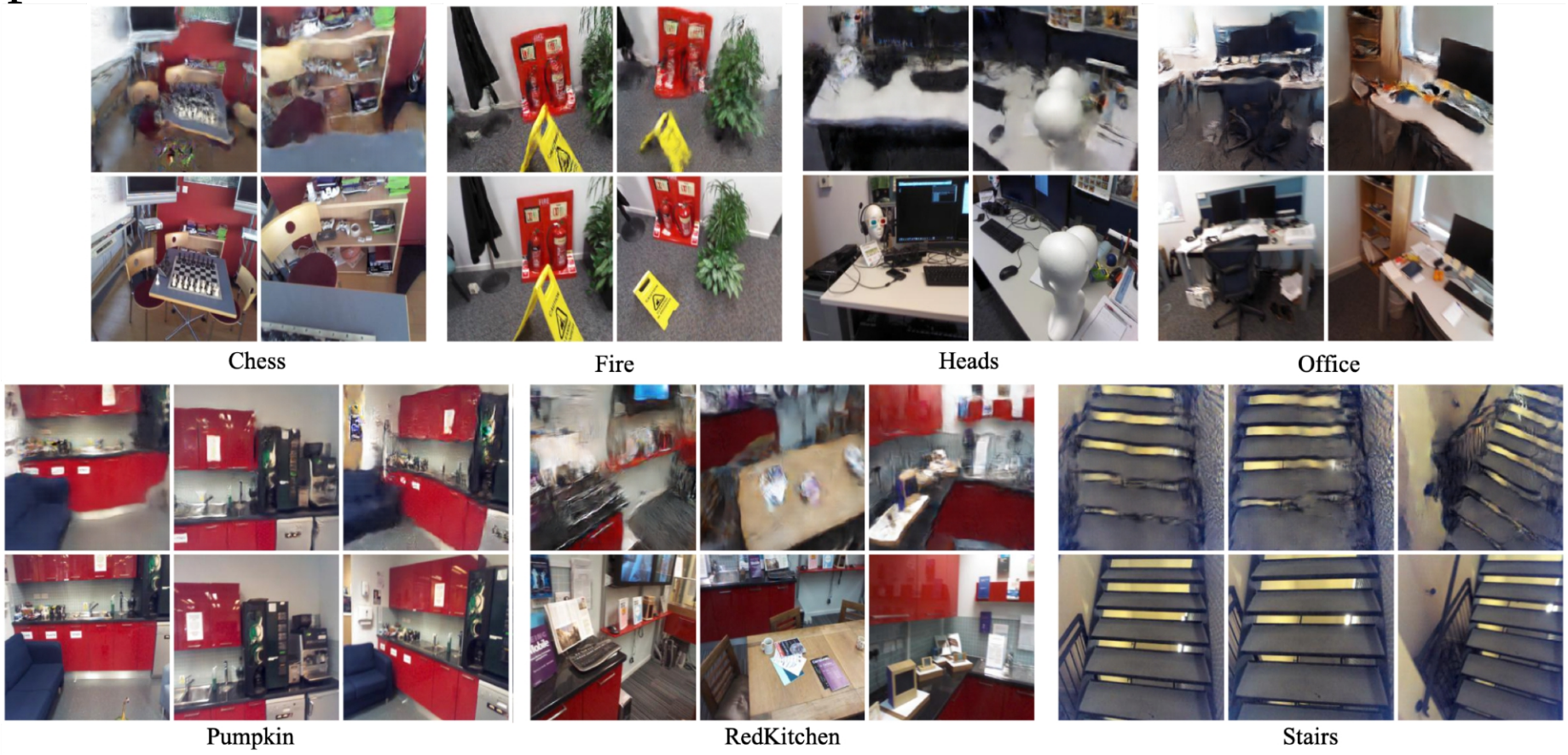
TABLE I

QUANTITATIVE EVALUATION OF THE SYNTHESIZED IMAGES QUALITY. THREE MEASURE METHODS WITH REFERENCE IMAGE: SSIM, PSNR, L_1 NORM AND ONE METHOD WITHOUT REFERENCE IMAGE: BRENNER. COARSE MEANS THE COARSE IMAGES GENERATED BY GENNET AND REFINED MEANS REFINED IMAGE BY REFINENET WITH OR WITHOUT PERCEPTUAL LOSS (PL).

[1] A. Kendall, M. Grimes, and R. Cipolla, “Posenet: A convolutional network for real-time 6-dof camera relocalization,” in *Proc. IEEE Int. Conf. Comp. Vis.*, 2015.

Results

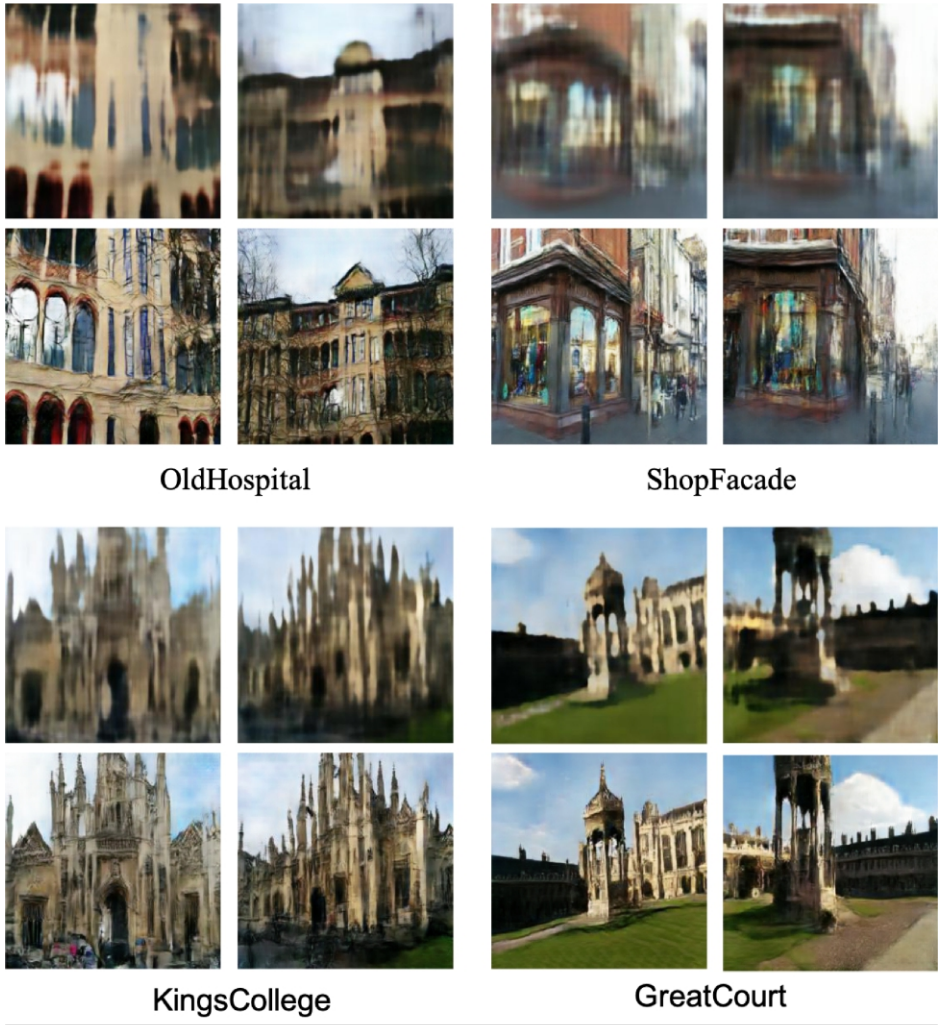
Examples on 7-Scenes [1]



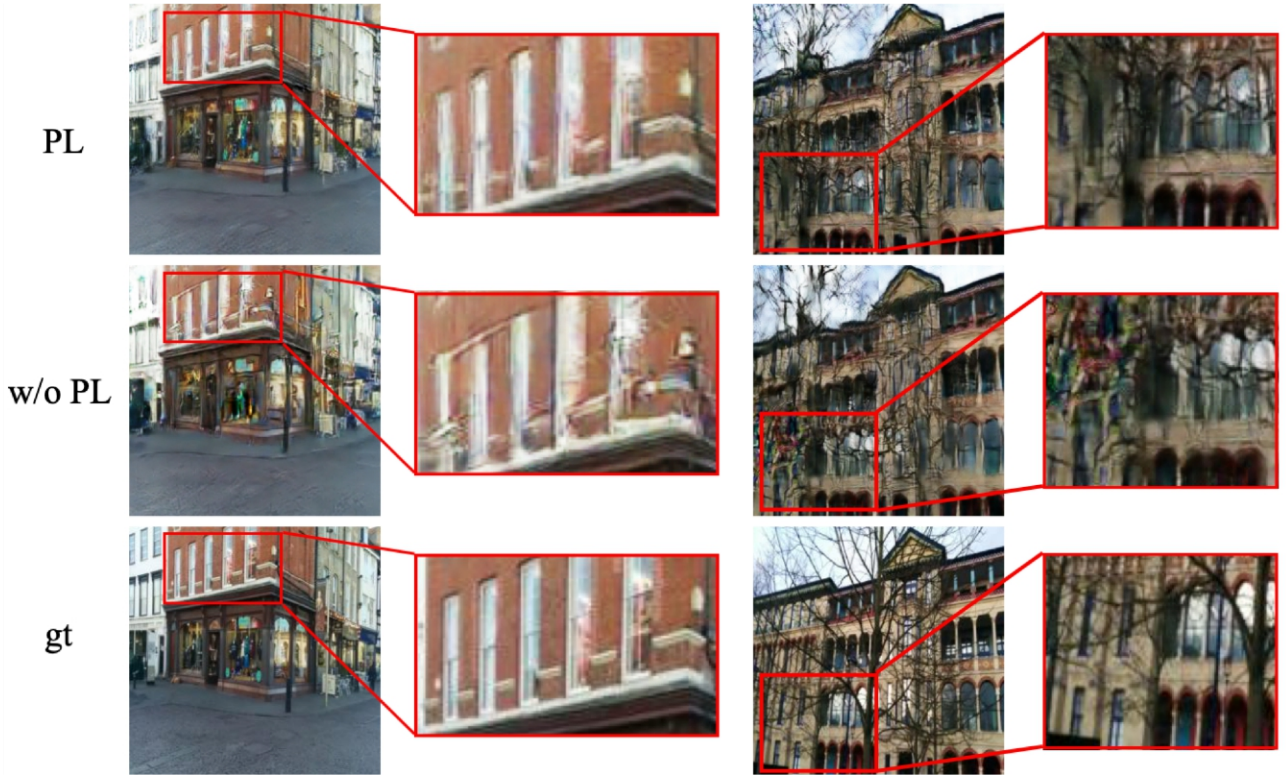
[1] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. W. Fitzgibbon, “Scene coordinate regression forests for camera relocaliza- tion in RGB-D images,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 2930–2937.

Ablation Studies

- Effect of RefineNet



- Effect of Perceptual Loss



Conclusion

- A new problem configuration of NVS: take only camera pose as input
- A two-stage training strategy which is consisted of two consecutive networks: GenNet and RefineNet, utilizing GAN and perceptual loss.
- Experiments show promising results in generating visually pleasant images
- Limitations: should be trained for each scene; distortion and geometric disalignment

Thank You!