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

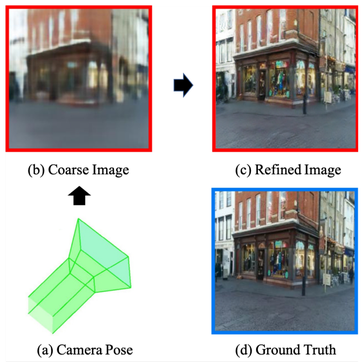
Xiang Guo, Bo Li, Yuchao Dai *, Tongxin Zhang, Hui Deng

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

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Introduction

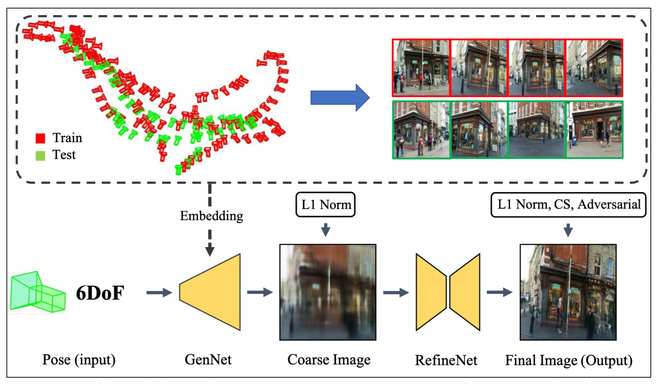
Novel view synthesis is a challenging problem in computer vision and robotics. Different from the existing works, which need the reference images or 3D models of the scene to generate images under novel views, we propose a novel paradigm to this problem. That is, we synthesize the novel view from only a 6-DoF camera pose directly. Although this setting is the most straightforward way, there are few works addressing it. While, our experiments demonstrate that, with a concise CNN, we could get a meaningful parametric model that could reconstruct the correct scenery images only from the 6-DoF pose. To this end, we propose a two-stage learning strategy, which consists of two consecutive CNNs: GenNet and RefineNet. GenNet generates a coarse image from a camera pose. RefineNet is a generative adversarial network that refines the coarse image.



Takes only the camera pose as input to predict the image under input view

Method

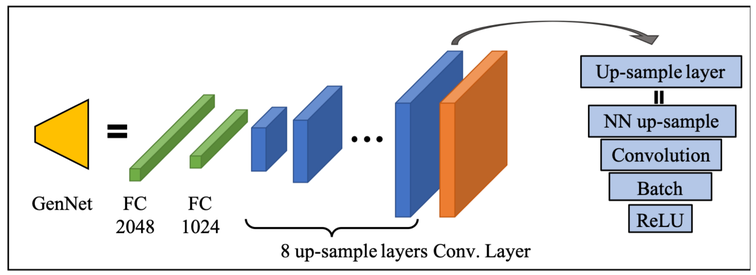
- **Overview**



Overview of the whole pipeline. A two-stage network consist of GenNet and RefineNet. The scene information is embedded into model parameters during training.

We propose a two stage framework consist of GenNet and RefineNet that connect in a sequence to synthesize an image from a camera pose. In the first stage, we use GenNet to get a coarse from pose P . In the second stage, we use RefineNet to refine the coarse image to get a fine detailed image.

- **GenNet**



Structure of GenNet. Use two fully connect layer to expand pose alone channel dimension

- **RefineNet**

We use U-Net structure, following pix2pix. The loss function includes L1 norm, Perceptual Loss and Adversarial Loss.

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

Results

- **On Cambridge Landmark**

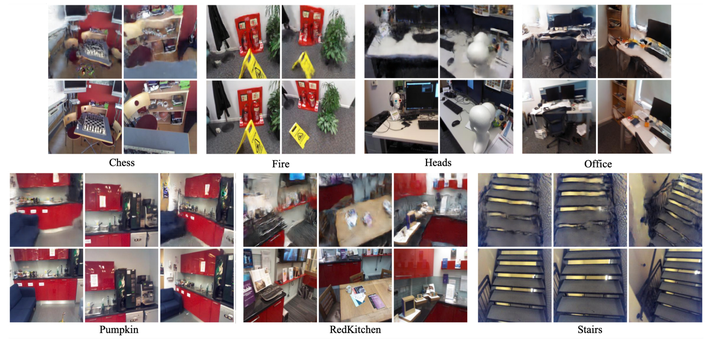
Examples of results on Cambridge Landmarks. For each scene, the first row contains synthesized images, and second row contains corresponding ground truth images

	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

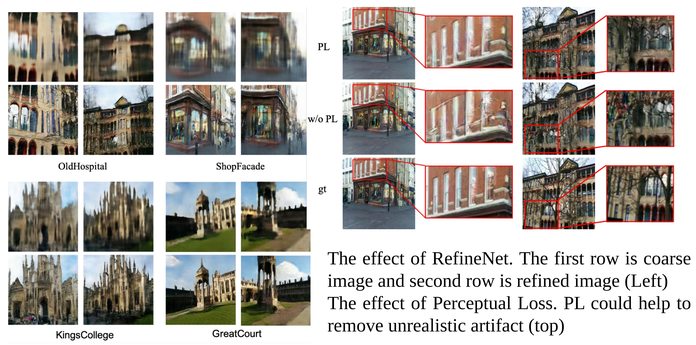
QUANTITATIVE EVALUATION OF THE SYNTHESIZED IMAGES QUALITY. THREE MEASURE METHODS WITH REFERENCE IMAGE: SSIM, PSNR, L_1 NORM AND ONE METHOD WITHOUT REFERENCE IMAGE: BRENNER. COARSE MEANS THE COARSE IMAGES GENERATED BY GENNET AND REFINED MEANS REFINED IMAGE BY REFINENET WITH OR WITHOUT PERCEPTUAL LOSS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

- **On 7-Scenes Dataset**



Examples of results on 7Scenes. For each scene, the first row contains synthesized images, and second row contains corresponding ground truth images

- **Ablation Studies**



The effect of RefineNet. The first row is coarse image and second row is refined image (Left)
The effect of Perceptual Loss. PL could help to remove unrealistic artifact (top)

Conclusion

We propose a new problem configuration of NVS: take only the camera pose as input. A two-stage framework consist of two consecutive networks: GenNet and RefineNet is used. Experiments show promising results in generating visually pleasant images. There are also limitations: need to train network for each scene; distortion in results and limited generalization ability.