Depth Assisted Full Resolution Network for Single Image-based View Synthesis

Xiaodong Cun
University of Macau
mb55411@umac.mo

Feng Xu
Tsinghua University
feng-xu@tsinghua.edu.cn

Chi-Man Pun*
University of Macau
cmpun@umac.mo

Hao Gao
Nanjing University of Posts and Telecommunications
tsgaohao@gmail.com

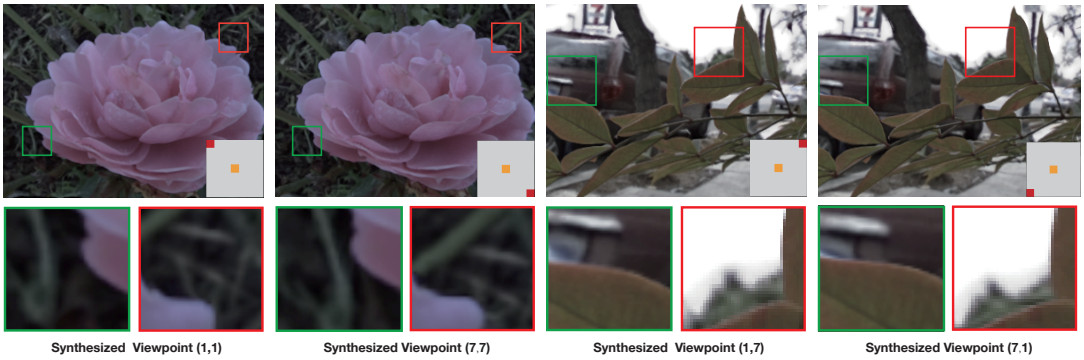

Figure 1: We propose a deep learning based method to synthesize surrounding novel views from the center view. We mark the coordinates of the center view as a yellow dot and the synthesized viewpoint as a red dot in the gray square, indicating the relative position of the viewpoints. Here we show our results at four extreme viewpoint positions. The zoomed-in regions contain both foreground and background whose relative positions are changed according to the changes in viewpoint.

CCS CONCEPTS

• **Computing methodologies** → **Image-based rendering**;

KEYWORDS

Depth based image rendering, Deep Learning, Light Field Image

ACM Reference Format:
Xiaodong Cun, Feng Xu, Chi-Man Pun, and Hao Gao. 2018. Depth Assisted Full Resolution Network for Single Image-based View Synthesis. In *Proceedings of SIGGRAPH '18 Posters.* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/3230744.3230789

1 INTRODUCTION

Synthesizing images of novel viewpoints is widely investigated in computer vision and graphics. Most works in this topic focus on using multi-view images to synthesize viewpoints in-between. In this paper, we consider extrapolation, and we take a step further to do extrapolation from one single input image. This task is very challenging for two major reasons. First, some parts of the scene may not be observed in the input viewpoint but are required for novel ones. Second, 3D information is lacking for single view input but is crucial to determine pixel movements between viewpoints. Although very challenging, we observe that human brains are always able to imagine novel viewpoints. The reason is that human brains have learned in our daily lives to understand the depth order of objects in a scene [Chen et al. 2016] and infer what the scene looks like when viewing from another viewpoint.

2 OUR APPROACH

We believe that for novel viewpoint synthesis, both global and local image features are important. And our key observation is that after modelling the process as two steps: depth prediction and depth-based image warping, the extraction of the two kinds of features can be decoupled. Depth estimation from a single image is ill-posed, so global high-level image features are required to tackle the problem. But given the depth, only local image warping is required to synthesize the final result, and local depth and color information are enough to determine the warping. Based on this observation, we focus on the two kinds of features in the two steps respectively.

As shown in Figure. 2, we explicitly estimate depth information from images by global high-level image features. As learning global features requires a large dataset to cover sufficient variations

*Cooresponding Author.

SIGGRAPH '18 Posters, August 12-16, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5817-0/18/08.
https://doi.org/10.1145/3230744.3230789

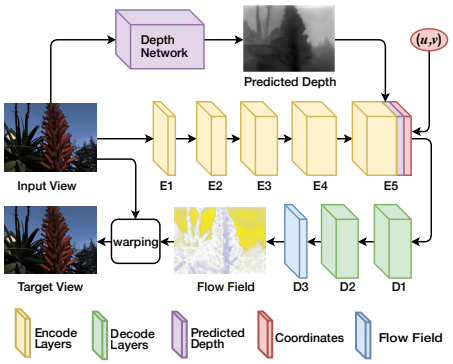

Figure 2: Network Structure.

and the current light field dataset is small, we leverage an existing 421k image dataset with labeled depth orders to pre-trained a *depth prediction network*[Chen et al. 2016].Then, with the good depth information, our view synthesis network is further trained to extract local features directly from the light field dataset. As local features do not have as much variations as global ones, current light field datasets are relatively sufficient. So we design a full resolution network to estimate the motion field from the input view to the target view with the learned features. Finally, A *warping* layer is used to not only warp the observed pixels to the desired positions but also hallucinate the missing pixels with recorded pixels. Following the idea of appearance flow [Zhou et al. 2016] , we apply flow based warping method for synthesizing the final image. For every pixel s in one novel view image, its pixel value can be expressed as: $I_q(s) = I_p[s+F_q(s)]$ where $F_q(s)$ is the two-dimensional flow which is the output of our neural network. Here, a backward warping is utilized to transform the input image to the novel view as the flow is defined at pixel s on the target view.

3 EXPERIMENTS

We test our method and the aforementioned four methods on all the 30 test images in the VS100 dataset and generate 48 novel viewpoints for each image. We compared many states-of-the-art related methods with our methods. Besides most relevant method [Srinivasan et al. 2017], we adapt the state-of-the-art methods for stereo pair synthesis[Xie et al. 2016], for synthesis with multi-view input[Kalantari et al. 2016] and for handling single object[Zhou et al. 2016], to fit our problem to perform the comparisons. We calculate three numeric metrics and represent the average values in Table 1. We can clearly see that our approach outperforms all the other related state-of-the-art methods.

4 APPLICATION

Synthesizing surrounding viewpoints from a single image has important applications such as free viewpoint rendering, synthetic apertures and refocusing by adding all the light field images together with different offsets, as shown in Figure 3. It can also be used to synthesize light fields on regular photos (by synthesizing dense viewpoints surrounding the input) and create VR applications by upsampling low ŋresolution light field imagery.

Table 1: Numerical comparison of our method and the state-of-art methods. A larger value indicates better quality for PSNR and SSIM, so does a less value for MAE.

	PSNR ↑	SSIM ↑	MAE ↓
[Srinivasan et al. 2017]	34.5788	0.8545	0.0285
[Kalantari et al. 2016]	34.1789	0.8483	0.0282
[Xie et al. 2016]	34.9809	0.8567	0.0232
[Zhou et al. 2016]	35.5367	0.8531	0.0237
Ours	**36.4401**	**0.8875**	**0.0202**

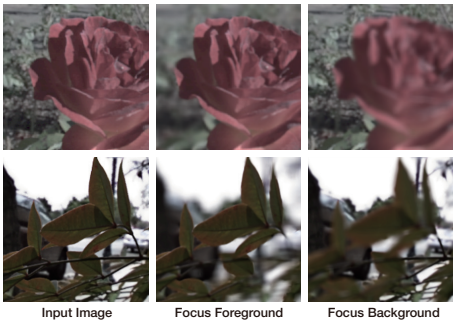

Figure 3: Some image refocusing examples. From left to right are the original input image, the image focusing on foreground and the image focusing on background.

5 CONCLUSION

In this paper, we propose a method to synthesize user-desired novel views from one single image. It is challenging and ill-posed, and difficult for current powerful deep learning techniques as there does not exist sufficient light field dataset for training. To tackle this problem, we first leverage a large image dataset with sparsely labeled depth orders to train a depth predictor. We demonstrate that combining the depth with only the local image features extracted by a specially designed full resolution network, novel view synthesis can be achieved on various input images.

ACKNOWLEDGMENTS

This work was supported in part by the Science and Technology Development Fund of Macau SAR under Grants 093/2014/A2 and 041/2017/A1.

REFERENCES

Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. 2016. Single-Image Depth Perception in the Wild. In *NIPS*. 730–738. http://papers.nips.cc/paper/6489-single-image-depth-perception-in-the-wild.pdf

Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. 2016. Learning-Based View Synthesis for Light Field Cameras. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2016)* 35, 6 (2016).

Pratul P. Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. 2017. Learning to Synthesize a 4D RGBD Light Field from a Single Image. *International Conference on Computer Vision (ICCV)* (2017).

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *ECCV*. Springer, 842–857.

Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. 2016. View Synthesis by Appearance Flow. In *ECCV*.