SceneNet: A generative model for generating scene from 6-DOF camera pose

Tuan-Anh Bui

February 25, 2018

Introduction

Given a set of images of a space, such as an office, which is captured by several cameras, how can infer a scene of that space when viewed from a particular point of view?

I noticed that despite there is plenty different approaches attack to this problem; we might see there is still not complete solution as we might hope. We might use some advanced methods of Structure-from-Motion (SfM) [16][14][15] to reconstruct the 3D model from the set of images then infer the scene we desire from 3D coordinate corresponding with 3D model. However, this approach cannot work on a low-depth object and requires to calibrate all camera, and the most important is that 3D model does not convey the appearance information. There are several other methods such as wrapping and combining images from nearby viewpoint [8][7][9]. However, those methods require two images taken from slightly different viewpoints.

In this project, I would like to propose a new end-to-end learning based generative model, that can generate a scene from a pose (position and orientation) directly. That model based on Generative Adversarial Networks [1] to learn a hidden transformation between two domains: pose domain and scene domain without explicitly understanding the distribution of two domains. To my best knowledge, this project will be the first approach in the field of view synthesis with an arbitrary pose.

Contribution to the Discipline

Generating a new synthesis image from a given images set is a highly active area in the field of computer vision and graphic. Flynn et al. [8] propose a novel learning-based model to predict new views from a given set of pose images on either side. The paper's architecture using two towers: selection tower and color tower, in which selection tower produces a probability map for each pixel to each depth that indicates the likelihood of each pixel having that depth. The color tower produces a full color for each pixel at each depth. However, the model requires a very specific preprocessing process that is *Plane Sweep Stereo (PSV)* reprojection. This technique reprojects two input images (from two different viewpoints to the target) to the different range of depth. Then the author uses the set of PSV as an input of learning model. Therefore, instead of developing a technique to learn a geometric mapping or directly enforce the epipolar constraint, the author using a typical model to learn an implicit input.

Another feasible approach is to use some advanced method of Structure From Motion (SFM) [16] algorithms or Simultaneous Localization And Mapping (SLAM) [14] [15] or to reconstruct a 3D model of space. Then use the view interpolation algorithms to infer the scene from arbitrary view to this 3D model, which including a variety of methods such as image correspondence and wrapping [19], and explicit shape and appearance estimation [20][22][21]. However, the disadvantage of this method is that it is not effective for low-level depth objects and that the 3d model does not carry appearance information such as color or texture.

The main contribution of this project is to propose an end-to-end model, that can learn directly not only the geometric transformation from each pose to each scene but also the appearance of the entire encoded space. If the project is successful, its capability can be used for many application such as virtual reality, teleconferencing [7], 3-dimensionalizing monocular film footage [9], video game [11], remote shopping.

Objectives of The Project

The achievement of this project are to:

- Deeply understanding and investigate the relationship between pose domain and scene domain, especially the hidden distribution under the geometric transformations.
- Provide the valuable analyses a possible of learning based model to learn the hidden geometric distribution
- Provide the practical analyses the effect of changing position and orientation to the performance of the model, also the capacity of the model to the capacity of embedding space.
- Provide a novel approach to embedding whole space into a model and generate a scene from arbitrary viewpoint and orientation.

Research Questions

New view synthesis is an extremely challenging, especially for the aim of this project is that generate a very-high dimensional space like scene images from a very-low dimensional space as poses (with only six dimensions including position and orientation in 3D space). Although, there have been a lot of view synthesis approaches that have been developed and have remarkable result [8] [9] [10], there are still difficult problems that need to be addressed in this project:

- Firstly, the most important aspect of the project is that there really exists a hidden distribution of the transformation between two domains: pose domain and scene domain. Based on the analysis and impressive result that Alex Kendall et al. have shown in [5], I strongly believe there is a hidden relationship between two domains, and that relationship can be learned through a deep network. So the first problem is transformed into another problem. How to design a deep network can learn a hidden relationship between two domains.

- Secondly, we can not completely learn the distribution without any constraint on geometric, especially homography. Thus, the second question is how to develop and integrate a geometric constraint into a deep network that helps to directly learn the hidden transformation between pose domain and scene domain.

- Thirdly, one of the most difficult problems of new view synthesis is lacking texture or occlusions. The traditional techniques such as multi-view stereo [9], image wrapping [7] often explicitly model the depth, color and occlusion components of each target pixel. The problems are that how to design an explicit model for occlusion into an end-to-end model. Can we just predict an unseen object by learning from common shapes and objects?

- Finally, how can we evaluate the performance of the generative model, especially the quality of the generated images? The most widely used full-reference image quality and distortion methods are peak signal-to-noise ratio (PSNR) and mean squared error (MSE), which do not correlate well with perceived quality [24][25]. Therefore the need to find a good method to assess the image quality to approximate the human perceptual.

Theoretical Framework and Methods

In order to achieve the goal in this project, I have several plans in my mind:

- Firstly, Generative Adversarial Networks [1] have been being the dominant approach for learning generative models. The key idea behind the huge success of GANs is that the model does not need to learn explicitly the data distribution, they just implicitly by doing the minimax game between generator and discriminator. Therefore, regardings to transfer learning between two domain, pose domain and scene domain, I will take time to investigate in detail with some state of the art GANs, e.g., [2] [4]

- Secondly, homography plays an important role in this problem, the good design of geometric constraint will help the model to learn more accurately. Therefore, I will take the time to familiarize myself with the problems involved in homography and the approaches related to homography.

- There are some steps to accomplish the fourth research question regarding the image quality assessment problem. The first step is to investigate and understand in detail some well-known image quality assessment methods, e.g., PSNR, MSSSIM [23], RECO [17], VIFP [18]. Then, do the small experiment to evaluate this method on the standard dataset to compare the difference, about a human assistant to correction. In my best knowledge, this project is the first approach. Therefore, I need to carefully investigate the good method

Baseline result

In order to assess the feasibility of the project, in other words, to answer the first research question of whether deep learning can learn the transformation from pose domain to scene domain, I have done the baseline experiment. In that experiment, I use the domain transfer model as described in [6], combined with the conditional GAN [3], training and

evaluating on two separate sets (training set and evaluated set) of 7Scenes dataset [26]. You can follow the video results in the link below (training process, testing process). In that experiment, I used the MSSSIM method [23] to evaluate the quality of the generated image. The results are quite limited as some scene cannot be synthesized, or the image is still blurred and noising. However, the results show that some of the capabilities of my approach.

References

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. "Generative Adversarial Networks", NIPS, 2014.

[2] Martin Arjovsky, Soumith Chintala and Lon Bottou. Wasserstein GAN, 2017; arXiv:1701.07875.

[3] Mehdi Mirza and Simon Osindero. "Conditional Generative Adversarial Nets", arXiv:1411.1784.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, arXiv:1703.10593.

[5] Alex Kendall, Matthew Grimes and Roberto Cipolla "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization." Proceedings of the International Conference on Computer Vision (ICCV), 2015.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks", CVPR 2017.

[7] A. Criminisi, A. Blake, C. Rother, J. Shotton, and P. Torr. "Efficient Dense Stereo with Occlusions for New View-Synthesis by Four-State Dynamic Programming", IJCV, 2007.

[8] John Flynn, Ivan Neulander, James Philbin and Noah Snavely. "DeepStereo: Learning to Predict New Views from the World's Imagery", CVPR, 2016.

[9] O. J. Woodford, I. D. Reid, P. H. S. Torr, A. W. Fitzgibbon. "On New View Synthesis Using Multiview Stereo", BMVC 2007.

[10] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, Alexei A. Efros. "View Synthesis by Appearance Flow", ECCV, 2016.

[11] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, Alexander C. Berg "Transformation-Grounded Image Generation Network for Novel 3D View Synthesis", CVPR, 2017.

[12] William Lotter, Gabriel Kreiman and David Cox. Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning, 2016; arXiv:1605.08104.

[13] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars and Luc Van Gool. "Pose Guided Person Image Generation", NIPS, 2017.

[14] Keisuke Tateno, Federico Tombari, Iro Laina and Nassir Navab. "CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction", CVPR, 2017.

[15] J. Engel and T. Schöps and D. Cremers "LSD-SLAM: Large-Scale Direct Monocular SLAM", ECCV, 2014.

[16] ChangchangWu. "Towards Linear-time Incremental Structure from Motion" 3DV, 2013.

[17] V. Baroncini, L. Capodiferro, E. D. Di Claudio, G. Jacovitti, "The Polar Edge Coherence: A Quasi blind metric for video quality assessment" EUSIPCO, 2009.

[18] Hamid R. Sheikh, Alan C. Bovik. "A visual information fidelity approach to video quality assessment", 2005.

[19] S. M.Seitz and C. R. Dyer. "View morphing", SIGGRAPH, 1996.

[20] S. Vedula, P. Rander, R. Collins, and T. Kanade. "Three-dimensional scene flow" IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(3), March 2005.

[21] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. "High quality video view interpolation using a layer representation", SIGGAPH, 2004.

[22] Q. Shan, R. Adams, B. Curless, Y. Furukawa, and S. M. Seitz. "The visual turing test for scene reconstruction" In Proc. 3D Vision, 2013.

[23] Zhou Wang, Eero P. Simoncelli and Alan C. Bovik. "Multi-Scale Structural Similarity for Image Quality Assessment" In IEEE Asilomar Conference on Signals, System and Computers, 2003.

[24] A. M. Eskicioglu and P. S. Fisher. "Image quality measures and their performance" IEEE Trans. Communications, vol. 43, pp. 29592965, Dec. 1995.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image quality assessment: From error measurement to structural similarity" IEEE Trans. Image Processing, vol. 13, Jan. 2004.

[26] Ben Glocker, Shahram Izadi, Jamie Shotton, Antonio Criminis. "Real-Time RGB-D Camera Relocalization" International Symposium on Mixed and Augmented Reality (ISMAR) IEEE October 1, 2013.