Advance in 3D visions - Final Report

Learning to Render Novel Views from Wide-Baseline Stereo Pairs

Group 4 劉冠宏311551058 周睦鈞 311553060 黃柏叡311553015

I. Introduction

Given a set of input images of a scene, the task of novel view synthesis is to produce plausible novel views, which are images of the same scene from viewpoints that do not appear in the inputs. Since NeRF was proposed, novel view synthesis has been one of the most active fields of research.

In the work "Learning to Render Novel Views from Wide-Baseline Stereo Pairs" novel views can be rendered with as sparse as a wide-baseline images pair. However, in most follow-up novel view synthesis works of NeRF, the input views need to be dense enough or each 3D point in the scene needs to be observed in multiple input views. Although some approaches have attempted to synthesize novel views with sparse observation as well, they are constrained in object-level scenes. In this work, reasonable novel views can be synthesized with only two views where plenty of 3D points are only observed in a single view.

Their method consists of three stages. First, the multi-view ViT encoder processes input images with self-attention across two views to obtain features for each view. Then, the image-centric epipolar line sampling strategy is proposed to efficiently sample the features to be used in rendering. Finally, the novel view rendering is achieved with a light-weight cross-attention renderer. Although the rendered quality is not comparable with methods using dense views input, it outperformed existing sparse view methods with the same settings.

We are interested in novel view synthesis, especially in the sparse views setting, so we chose this paper as our reproducibility challenge topic, tested their methods in different experiment settings and with data prepared by ourselves.

II. Background

Since the introduction of NeRF in 2020, novel view synthesis has become one of the most popular fields of research in computer vision. NeRF constructs a neural radiance field implicitly through optimizing a multi-layer perceptron (MLP) with a set of input images and their camera poses. By querying the MLP with coordinates of 3D points and viewing directions in world space, the RGB color and density value used in volume rendering can be obtained.

Although NeRF has set the new state-of-the-art in several novel view synthesis datasets, there is still significant room for improvement. The main restrictions of NeRF include the

need for dense input views, accurate camera poses, and time-consuming per-scene optimization. One research direction is to reduce the requirement of dense input views. pixelNeRF is one of the methods that can synthesize novel views with as few as a single view. However, these works are limited to object-level scenes or do not scale well to complex real-world scenes. In this challenging problem setting, the work "Learning to Render Novel Views from Wide-Baseline Stereo Pairs" takes a wide-baseline stereo image pair using proposed image-centric epipolar line sampling that enables them to fully utilize image features efficiently, rendering outstanding results compared to existing methods.

III. Method

They use a multiview encoder to compute pixel-aligned features, and a cross attention-based renderer to transform the features into novel view renderings, see Figure 2 for an overview.

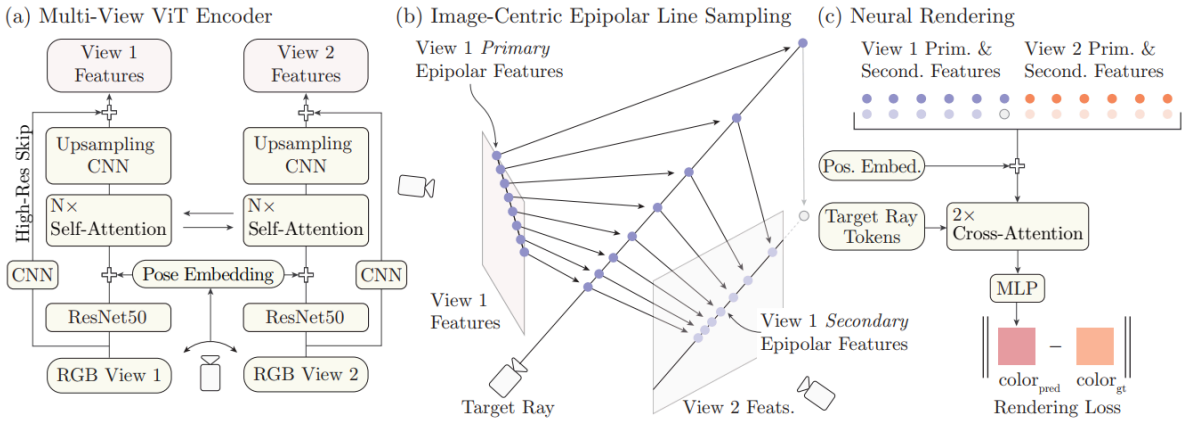

Figure 2. **Method Overview.** (a) Given context images from different viewpoints, a multi-view encoder extracts pixel-aligned features, leveraging attention across the images and their corresponding camera pose embeddings. (b) Given a target ray, in each context view, we sample *primary* features along the epipolar line equidistant in pixel space. We then project the corresponding 3D points onto the other views and sample corresponding *secondary* epipolar line features, where out-of-bounds features are set to zero. (c) We render the target ray by performing cross-attention over the set of all primary and secondary epipolar line features from all views.

i. Multiview Feature Encoding

In prior work, pixel-aligned features are obtained by separately encoding each image via a vision transformer or CNN. However, it led to artifacts in renderings observing boundary regions between context images. The authors hypothesize that separate encoding of images leads to inconsistent geometry reconstruction across context images. Thus, they proposed a multi-view encoder to solve the problem.

The architecture contains the following part. First, ResNet50 CNN extracts features for two input images independently. Second, to each feature, they add positional embedding and a camera pose embedding. Third, a vision transformer performs self-attention across all tokens across both images. Fourth, a fusion CNN used to upsample to get spatial feature maps. Directly using these spatial feature maps for novel view synthesis leads to blurry results, due to the loss of high-frequency texture information. Therefore, They concatenate these features with high-resolution image features obtained from a shallow CNN.

ii. Epipolar Line Sampling and Feature Matching

Instead of using volume rendering, they proposed a sampling strategy by using an epipolar line, which is simple but more effective. For each target ray, the epipolar line can be found on the feature map. They uniformly sample N pixel coordinates along the line segment of the epipolar line within the image boundaries, where they call **primary features**. To enable the renderer be able to choose a certain pixel-aligned feature or not. The **depth value** on the feature map is important, and it can be computed via triangulation. In order to refine geometric information to get more consistent results. They proposed a feature matching module. First, they project a primary feature to a 3D point, and then project this 3D point onto the other feature map to retrieve a corresponding feature, which we refer to as a **secondary feature.** In the end, we can get a set that contains depth, primary feature and secondary feature $\{(d, \mathbf{f}, \hat{\mathbf{f}})_k)_{k=1}^{2N}$.

iii. Differentiable Rendering via Cross-Attention

To render the target ray, it remains to map the set of epipolar line samples $\{(d, \mathbf{f}, \hat{\mathbf{f}})_k)_{k=1}^{2N}$ to a color value. They proposed a cross-attention decoder. First, they embed the target ray origin, target ray direction, depth with respect to the target ray origin, and context camera ray direction. Second, perform two rounds of cross-attention to obtain a final feature embedding. Finally, decode into color via a small MLP.

vi. Training and Losses

The loss function consists of two terms. LPIPS perceptual and regularization. LPIPS loss is used to compute the similarity of two images. Regularization loss is for better multi-view consistency. For better generalization, they perform some data augmentations during the training procedure. They center crop and scale the input and target images, which leads to transformation in the intrinsics of the camera. They also flip the images which leads to transformation of the extrinsics.

$$\mathcal{L} = \mathcal{L}_{\text{img}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} \,. \tag{2}$$

$$\mathcal{L}_{\text{img}} = ||R - G||_1 + \lambda_{\text{LPIPS}} \mathcal{L}_{\text{LPIPS}}(R, G) \,, \tag{3}$$

$$\mathcal{L}_{\text{reg}} = \sum_{(u,v)} \sum_{(u'v') \in \mathcal{N}(u,v)} ((e(u,v) - e(u',v'))^2 \,. \tag{4}$$

IV. Experiment

i. Dataset

They train and evaluate our approach on RealEstate10k, a large dataset of indoor and outdoor scenes, and ACID, a large dataset of outdoor scenes. They use 67477 scenes for training and 7289 scenes for testing for RealEstate10k, and 11075 scenes for training and 1972 scenes for testing for ACID.

They train their method on images at 256 x 256 resolution and evaluate methods on their ability to reconstruct intermediate views in test scenes.

ii. Qualitative Results

Comparative Rendering Results on RealEstate10k.

Their approach can render novel views of indoor scenes with substantial occlusion with high fidelity using a wide-baseline input image pair, outperforming all baseline. Note that many points of the 3D scene are only observed in a single image in such inputs.

Comparative Results on ACID.

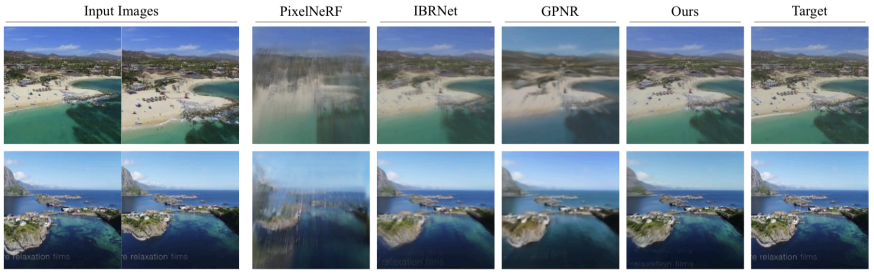

Figure 5. **Comparative Results on ACID.** Our approach is able to render novels views with higher quality than all baselines.

iii. Quantitative Results & Ablation study

Method	LPIPS ↓	SSIM ↑	PSNR ↑	MSE ↓
pixelNeRF [59]	0.591	0.460	13.91	0.0440
IBRNet [54]	0.532	0.484	15.99	0.0280
GPNR [48]	0.459	0.748	18.55	0.0165
Ours	**0.262**	**0.839**	**21.38**	**0.0110**

Table 1. **Novel view rendering performance on RealEstate10K.** Our method outperforms all baselines on all metrics.

Method	LPIPS ↓	SSIM ↑	PSNR ↑	MSE ↓
pixelNeRF [59]	0.628	0.464	16.48	0.0275
IBRNet [54]	0.385	0.513	19.24	0.0167
GPNR [48]	0.558	0.719	17.57	0.0218
Ours	**0.364**	**0.781**	**23.63**	**0.0074**

Table 2. **Novel view rendering performance on ACID.** Our method outperforms all baselines on all metrics.

Models	LPIPS↓	SSIM↑	PSNR↑	MSE↓
Base Model	0.452	0.735	18.11	0.0201
+ 2D Sampling	0.428	0.762	19.02	0.0159
+ Cross Correspondence	0.415	0.766	19.52	0.0142
+ Multiview Encoder	0.361	0.794	20.43	0.0132
+ Regularization Loss	0.358	0.808	19.84	0.0139
+ Data Aug	0.262	0.839	21.38	0.0110

Table 3. **Ablations.** All components of our proposed method are essential for high-quality novel view synthesis.

Quantitatively evaluate their approach and baselines in Table1 and Table2. Finding that their approach substantially outperformed each compared baseline in terms of all our metrics.

V. Reproduce

i. Training and Evaluation

To reproduce the paper, We follow the source github to build up the environment. We download and unzip the release weights from source github, and evaluate their model. For evaluation and training, we use the subset of realestate10K, which can be downloaded through source github. Because the dataset is really large, we just use a part of it to speed up the process.

The reproducibility of evaluation and training requires a GPU of more than 11GB. We tried to reproduce using 2080 Ti 11GB but it failed. We reproduce successfully with A5000 24GB.

We use the commands which are referred to the source github to start our training and evaluation.

ii. Experiments

Beside the training and evaluation, we also try some of our images which are different from the images in paper. We want to know how the perspective influences the novel view synthesis.

We use the two images below as the input.

Output:
https://drive.google.com/file/d/1fqW6Lz3tO7VxFY64BgsrJtI24gjzI016/view?usp=share_link

We use the other two of our input images to render the novel view video. However, the view of the image is more different.

Output:
https://drive.google.com/file/d/1d8jOx2ntUY_C8-CLe6jpr-dCLlVrCX9i/view?usp=share_link

Compared with the first experiment, the result is worse. There are more artifacts in the rendering video.

VI. Reference

[1] Du, Yilun, et al. "Learning to Render Novel Views from Wide-Baseline Stereo Pairs." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[2] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM* 65.1 (2021): 99-106.

[3] Yu, Alex, et al. "pixelnerf: Neural radiance fields from one or few images." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.