

Ultra-Sparse Viewpoints Novel View Synthesis via Global Feature-Driven Spatial Structure Comprehension *

Qijun He¹ [0009-0006-8413-703X], Jingfu Yan¹ [0009-0005-8767-4749], Jiahui Li¹ [0009-0003-2000-9432],
and Yifeng Li¹ (✉) [0000-0003-4798-3211]

¹ College of Computer and Information Engineering, Nanjing Tech University, Nanjing,
Jiangsu 211816, China
lyffz2616@163.com

Abstract. Our research focuses on solving the numerous artifacts and geometric distortion problems encountered when synthesizing new views from extremely sparse input views. We find that enhancing global features under such sparse conditions helps the network better understand the spatial relationships of the scene, thereby improving rendering quality. Our method is divided into two main parts: In the first part, our method emphasizes global feature extraction. In the case of sparse views, we use more global features to make up for the shortcomings of other view features and utilize these features to provide the network with a comprehensive grasp of the overall layout and structure of the scene. The second part uses a mechanism similar to the visual transformer to fuse features from each input view, and uses ViT to enhance the model's understanding of different spatial relationship features to solve the problem of multi-viewpoint geometric consistency. Experiments show that when tested on the most popular real-scenario forward datasets and synthetic datasets, our approach exhibits state-of-the-art performance and demonstrates richer performance compared to previous excellent work on synthesizing new views. details and a more complete outline structure.

Keywords: Transformer · ViT · NeRF · Sparse Views

1 Introduction

Before our approach, there were numerous strategies for reconstructing new views from sparse inputs, including RegNeRF [1], DDP-NeRF [2], DS-NeRF[3], ViewFormer [4], and the recently introduced GBT [5], DiffusioNeRF [6], ViP-NeRF [7], and FreeNeRF [8], each providing robust solutions to the sparse view reconstruction dilemma. For instance, FreeNeRF [8] leverages the regularization of the visible spectrum to mitigate smooth transitions, thereby incrementally enriching the radiance field with high-frequency details. RegNeRF [1], a patch-based regularizer, enhances geometric integrity by minimizing floating artifacts. However, these methods predominantly focus on addressing sparse perspective challenges and remain confined to rendering new

* Q. He and J. Yan—These authors contributed equally to this work.

viewpoints within isolated scenes. Models like MVSNerF [9], pixelNeRF [10], SRF [11], and IBRNet [12] surmount these limitations. While these models exhibit generalizability to new scenes, their performance falters under extreme view sparsity; the feature extraction from limited viewpoints becomes inadequate, causing structural distortions, marked degradation in image fidelity, and pronounced artifact and “floater” manifestations.

To address these challenges, our method is trained by sparsely sampling the input views around the target pose and incorporating them into an end-to-end network framework. This strategy enables our model to generalize to new data domains, thus overcoming the limitation of NeRF [13] requiring individual scene optimization. In the feature extraction stage of the input view, we integrated the Transformer encoder to broaden the receptive field and enhance global feature extraction, so that more global features can serve as features of the corresponding structure of the unknown perspective, making up for the difficulties caused by sparse perspectives. At the same time, the recent breakthroughs in image processing achieved by the Vision Transformer (ViT [14]) model, especially its powerful capability in feature fusion within a single image, motivate us to leverage the power of ViT to perform multi-view fusion tasks. In particular, we utilize ViT to merge and merge features from each viewpoint, solving the complex problem of multi-view geometric and photometric consistency. Our extensive experimental studies demonstrate that our approach outperforms previous methods. Essentially, our contribution is:

1. To make up for the difficulties caused by view sparsity, we use advanced encoder-decoder and Transformer hybrid architecture to enhance global features, so that more global features can be used as features of unknown view corresponding structures.
2. We integrate a Vision Transformer (ViT) component to amalgamate features from input views with varying poses, thereby augmenting the geometric and photometric coherence.
3. We successfully synthesize clearer novel views from sparse input data, markedly diminishing the incidences of artifacts and geometric distortions.

2 Related work

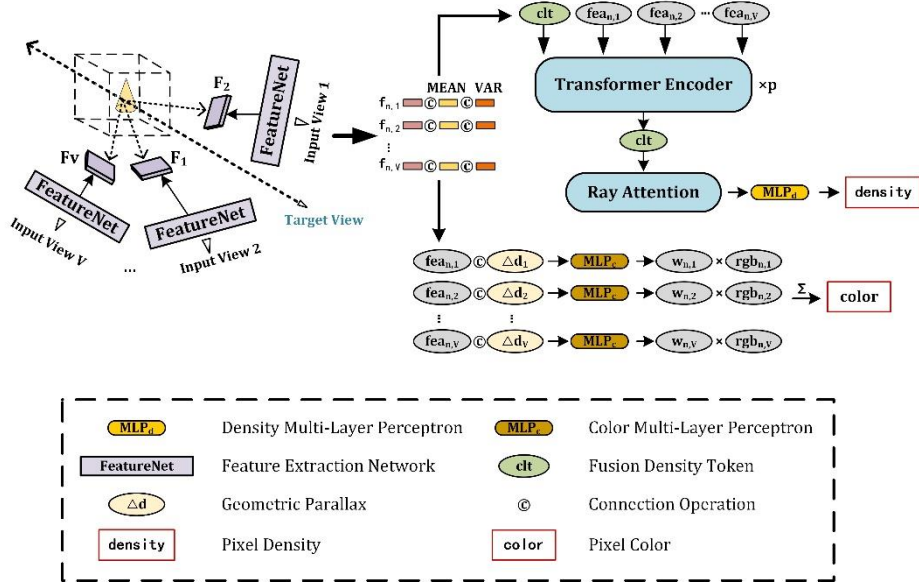
More recent work, like Neural Radiance Field (NeRF [13]), conceptualizes a scene by representing each point in space as a compact, continuous 5D function. This method uses a multilayer perceptron (MLP) to model the scene’s radiance field, detailing the color and direction of each point. However, training for a single scene requires hundreds of nearby views and extensive optimization for each scene. To address the issue of numerous input views, subsequent research such as DS-NeRF [3] proposes solutions like deriving sparse depth from structure from motion and integrating depth regularization into the loss function. Yet, these approaches do not eliminate the need for per-scene optimization. Meanwhile, the recent MVSNerF [9] has also shown excellent performance. However, creating 3D features from planar scanned volumes typically demands significant memory resources. While 2D feature extraction is more memory-efficient,

reconstructing new views with limited input frequently results in severe degradation of synthetic image quality due to the insufficiency of 2D features.

3 Method

Our approach primarily deduces the density and color of 3D points by extracting 2D features from the input views. This process unfolds in three pivotal stages: initially, we concentrate on extracting 2D features from the input views; subsequent is the phase of feature fusion and geometric reasoning; the final part entails the application of neural volumetric rendering. The overall model is shown in **Fig. 1**.

Fig. 1. Overview of our approach



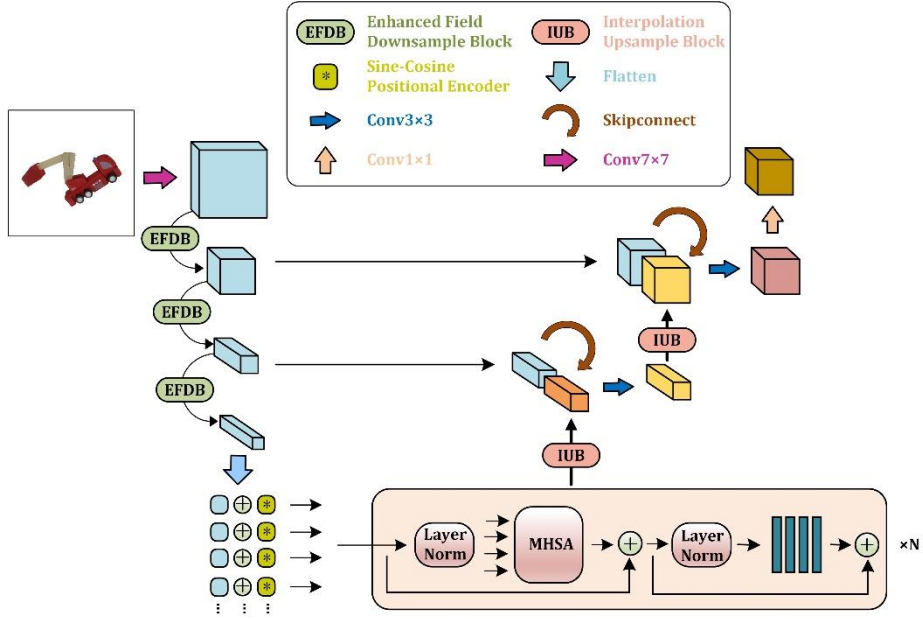
3.1 Feature Extraction

As shown in **Fig. 2**. Our feature extraction model is based on a hybrid model of encoder-decoder and Transformer to extract 2D features $\{F_i\}_{i=1}^V \in R^{\frac{H_i}{4} \times \frac{W_i}{4} \times d}$ from all source views $\{I_i\}_{i=1}^V \in R^{H_i \times W_i \times 3}$. We do this by projecting the 3D points $pts \in R^3$ on the rendering rays to all input views through the camera parameters, and in order to maintain the continuity of features, we use bilinear interpolation [15] to extract the features $\{f_v\}_{v=1}^V \in R^d$ corresponding to the projected points, where V represents the number of input views and d represents the feature dimension:

$$\{F_i\}_{i=1}^V = M(\{I_i\}_{i=1}^V), \quad (1)$$

Where $M(\cdot)$ represents the hybrid model of the encoder-decoder and Transformer.

Fig. 2. Overview of the feature extraction network: During the encoder stage, the convolutional layers execute downsampling operations to capture local deep features, such as textures and details. Subsequently, the network obtains positional information through a sine-cosine positional encoder. This information is then fed into N layers of Transformer Encoder modules, which enhance the global feature ratio, further augmenting the model’s understanding of the overall spatial relationships. In the decoder stage of the feature extraction network, the upsampling process employs bilinear interpolation. Additionally, each layer is connected via skip connections to the corresponding features of the encoder layer, providing rich local features to assist the network in more accurately depicting small objects and edge textures within the image.



3.2 Volume Density Prediction

We predict the bulk density of a 3D point (x, d) in two stages, where d denotes the 3D spatial ray direction, and x signifies a sampling point along that direction. Initially, our method consolidates the single-point features from all input views corresponding to the 3D point. Subsequently, to overcome the constraints associated with single-point features, we implement a multi-head attention mechanism. This mechanism is designed to focus on learning the intricate relational dynamics among density features from all sampled points along the identical ray direction.

We define all sampling points in the same ray direction as $\{x_n\}_{n=1}^N$, where N represents the number of sampling points in the ray direction, and each sampling point is mapped to the feature map $\{F_i\}_{i=1}^V$ corresponding to all input views to obtain a single point feature $\{f_{n,v}\}_{v=1}^V$. Next, we add global feature $MEAN(\{f_{n,v}\}_{v=1}^V)$ and feature dispersion $VAR(\{f_{n,v}\}_{v=1}^V)$ to avoid outliers and focus on important features.

Specifically, the global features and feature dispersion are respectively spliced to all single-point features to obtain $\{fea_{n,v}\}_{v=1}^V$:

$$\{fea_{n,v}\}_{v=1}^V = \langle \{f_{n,v}\}_{v=1}^V \mid MEAN\{f_{n,v}\}_{v=1}^V \mid VAR\{f_{n,v}\}_{v=1}^V \rangle, \quad (2)$$

Where $\langle \cdot \mid \cdot \mid \cdot \rangle$ represents the connection operation of feature dimensions, $\{fea_{n,v}\}_{v=1}^V$ represents the input feature after the single-point feature is spliced with global features and discreteness. Next, we use part of the architecture in ViT to complete the fusion of input feature $\{fea_{n,v}\}_{v=1}^V$. Since there is no positional correlation between our input features $\{fea_{n,v}\}_{v=1}^V$ and no Patch operation is needed to aggregate local features, we abandon the position encoding stage and Patch stage in ViT, and treat the input feature $\{fea_{n,v}\}_{v=1}^V$ as the input sequence in ViT. Within the ViT model, the features $\{fea_{n,v}\}_{v=1}^V$ of different input views learning cross-view correlation through self-attention mechanism. Next, we chose to add Class Token (clt) as the final fusion density token $\{fea'_n\}_{n=1}^N$, which can effectively balance the features from different views and reduce the bias caused by single view features and improve the overall consistency of the synthesized image. Through this method, the ViT model is able to generate a global feature representation that not only contains the information of each individual view, but also contains the interaction information between multiple views:

$$\{fea'_n\} = Trans^{\times p} \left(\left[\{fea_{n,v}\}_{v=1}^V, \{fea'_n\} \right] \right) \quad \forall n \in \{1, 2, \dots, N\}, \quad (3)$$

Where $[\cdot, \cdot]$ represents dimension splicing, $Trans^{\times p}$ represents p-layer Transformer Encoder.

We further regularize the fused density token using the Multihead Attention Mechanism (MHA) so that each sampling point can learn a global density feature on the ray:

$$\{f_n^\sigma\}_{n=1}^N = MHA(\{fea'_n\}_{n=1}^N). \quad (4)$$

Finally, we use a multilayer perceptron to reason about the density of a spatial point $\{\sigma_n\}_{n=1}^N$:

$$\{\sigma_n\}_{n=1}^N = MLP_d(\{f_n^\sigma\}_{n=1}^N), \quad (5)$$

Where $MLP_d(\cdot)$ represents a four-layer perceptron.

3.3 Color Prediction

For color prediction, we add relative parallax information to the input features as a way to learn the similarity between the novel view and the input view. The smaller the parallax, the more likely we consider the input view to be similar to the novel view, and the larger the corresponding input feature weights. The addition of parallax information ensures that information captured from different viewing angles is geometrically and color consistent during composition, resulting in images that look natural and are geometrically correct. Specifically, we use the geometric parallax $\{\Delta d_v\}_{v=1}^V$ between the

input view and the target view to splice with the input feature $\{fea_{n,v}\}_{v=1}^V$. Next, we predict the color weight of each input view mapping point by a multilayer perceptron:

$$w_{n,v} = softmax\left(MLP_c(\{fea_{n,v}\}_{v=1}^V \mid \{\Delta d_v\}_{v=1}^V)\right) \quad \forall n \in \{1, 2, \dots, N\}, \quad (6)$$

Where $MLP_c(\cdot)$ represents the multi-layer perceptron that infers color features, $softmax(\cdot)$ represents the normalized exponential function to predict color weights.

Next, the color weight is multiplied by the corresponding color and summed to obtain the color \hat{c}_n of the space point:

$$\hat{c}_n = \sum_{v=1}^V w_{n,v} rgb_{n,v} \quad \forall n \in \{1, 2, \dots, N\}, \quad (7)$$

Where $rgb_{n,v}$ represents the color of all input view mapping points corresponding to each spatial sampling point.

3.4 Volume Rendering

For volume rendering, we follow the NeRF [13] volume rendering principle by weighting the color of each point on the ray according to its volume density and transmission function along the ray direction, and then accumulating these colors to get the final color $\hat{C}(r)$ of the ray:

$$\hat{C}(r) = \sum_{n=1}^N T_n (1 - \exp(-\sigma_n)) \hat{c}_n, \quad (8)$$

$$T_n = \exp(-\sum_{j=1}^{n-1} \sigma_j), \quad (9)$$

Where T_n represents the cumulative transmittance of all points before the light passes through point n.

3.5 Loss Functions

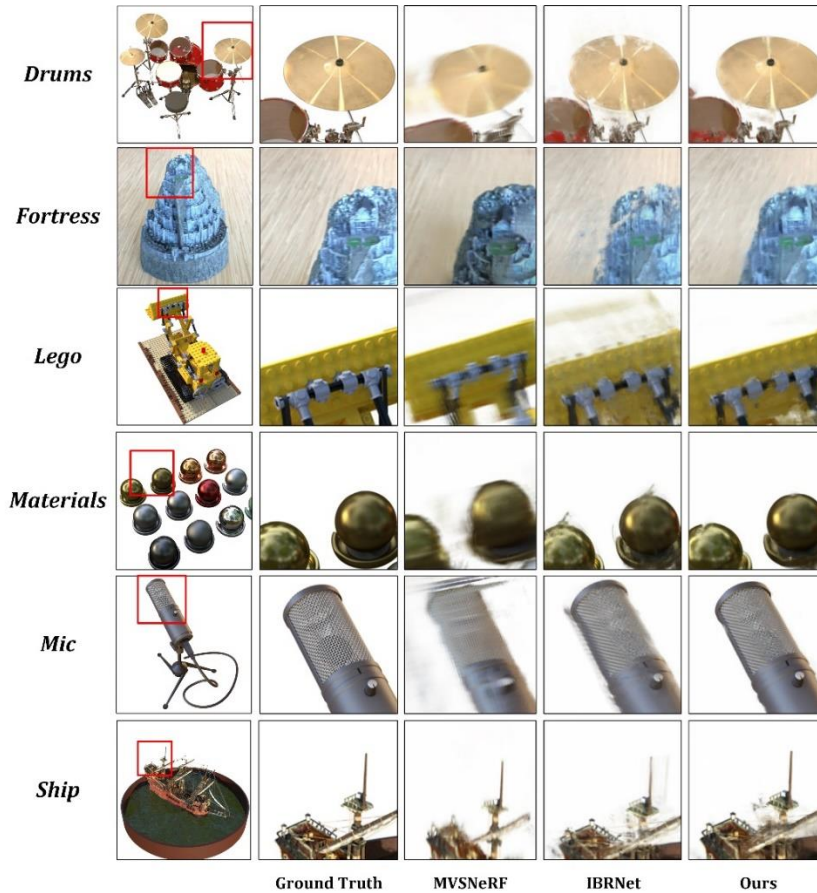
For the loss function, we employ the Mean Squared Error (MSE). Given our adoption of NeRF’s hierarchical volume sampling strategy, we compute the MSE for the ray rendering results at both coarse and fine sampling levels relative to the ground truth and subsequently aggregate these errors.

$$loss = \sum_{r \in R} \left[\|\hat{C}_c(r) - C(r)\|_2^2 + \|\hat{C}_f(r) - C(r)\|_2^2 \right], \quad (10)$$

where R represents the set of rays during training, $C(r)$ represents the true color, $\hat{C}_c(r)$ and $\hat{C}_f(r)$ represent the final rendered colors of coarse sampling and fine sampling respectively.

4 Experiments

Fig. 3. Qualitative comparison of methods on synthetic datasets[13] (Drums, Lego, Mic, Materials, and Ship) and the real forward datasets[18] (Fortress) demonstrates our proposed method's significant enhancement in scene detail retention, geometric appearance accuracy, and reduction in visual artifacts (floaters). Conversely, IBRNet exhibits pronounced artifacts and blurry boundaries. For instance, in processing the Fortress scene, our model effectively maintains structural continuity and integrity, in stark contrast to the image fracturing observed with IBRNet. Additionally, when addressing boundary details, such as the edge processing in Mic and Ship scenes, our approach substantially minimizes blur artifacts, thereby preserving the image's overall visual quality.



Dataset. Our training dataset consists of three parts in total. The first part of the dataset utilizes Google Scanned Objects [16], which renders images centered on the object. We use the Spaces dataset [17] as our second part. For the final part, we used 67 real scenes captured by mobile phones, collected from the IBRNet collection [12]. Among these, the Google Scanned Objects Dataset [16] contains 1023 models. The Spaces dataset

[17] includes 100 scenes captured by 16 cameras. As for the dataset of real scenes captured by mobile phones [12], each scene comprises 20-60 images taken by the front-facing camera, roughly distributed on a 2D grid [18]. Our evaluation dataset employs synthetic renderings of objects and real images of complex scenes. For the synthetic rendering dataset of objects, we utilize the NeRF synthetic dataset [13], which includes 100 training views and 200 test views at 800×800 resolution, sampled on the upper hemisphere or the entire sphere. For real datasets, we use roughly forward images from NeRF [18], with each scene consisting of 20 to 62 images captured at a 1008×756 resolution.

4.1 Experimental Results

Our evaluation of the model is divided into two parts. Initially, our model was compared against existing generalizable NeRF models such as IBRNet [12], GeoNeRF [19], and MVSNerF [9] without per-scene optimization; the comparison details are shown in **Fig. 3**. Subsequently, under the context of per-scene optimization, comparisons are drawn between our model and other models including DietNeRF [20], RegNeRF [1], DS-NeRF [3], DDP-NeRF [2], and ViP-NeRF [7].

In the field of visual rendering and reconstruction, quantitative evaluation metrics such as Peak Signal-to-Noise Ratio (PSNR [21]), Structural Similarity Index (SSIM [22]), and Learned Perceptual Image Patch Similarity (LPIPS [23]) play a pivotal role in gauging model performance. In our study, we conducted a quantitative comparison using these metrics on a synthetic dataset, as depicted in **Table 1**. The findings demonstrate that our proposed model surpasses existing generalizable models in processing input images with identical viewing angles and quantities, exhibiting superior image reconstruction precision and visual quality. Notably, our model exhibits outstanding perceptual consistency on the LPIPS metric, thereby enhancing the image content’s realism. **Table 2** presents a comparative evaluation on a real-world dataset, illustrating our model’s exceptional capability to restore image details and overall structure, particularly excelling in high-contrast areas and preserving luminance consistency.

Table 1. Performance on SYNTHETIC DATASETS [13].

Method	Settings	Synthetic Data [13]		
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
IBRNet [12]	No per-scene Opt	21.38	0.850	<u>0.195</u>
MVSNerF [9]		<u>22.60</u>	<u>0.867</u>	0.238
GeoNeRF[19]		16.09	0.653	0.390
Ours		22.96	0.876	0.155

Table 2. Evaluation on Real Forward-Facing Datasets [18].

Method	Settings	Real Forward-Facing Data[18]		
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
IBRNet [12]		21.22	<u>0.681</u>	<u>0.339</u>
MVSNeRF [9]	No per-scene Opt	17.08	0.574	0.463
GeoNeRF[19]		17.41	0.415	0.528
Ours		<u>21.13</u>	0.691	0.322

Table 3. Model convergence experiments on SYNTHETIC DATASETS [13].

Method	Settings	Synthetic Data [13]		
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Our _{Sno-ft}		22.96	0.876	0.155
Our _{10k}	No per-scene Opt	24.71	0.899	0.119
Our _{30k}		25.04	0.901	0.114

Table 4. Comparison of per-scene rendering work with sparse input.

Method	Number of Views	Real Forward-Facing Data[18]		
		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
DietNeRF[20]		11.89	0.320	0.726
RegNeRF[1]		16.90	0.487	0.440
DS-NeRF [3]	No per-scene Opt	17.06	0.506	0.454
DDP-NeRF [2]		<u>17.21</u>	<u>0.537</u>	0.422
ViP-NeRF [7]		16.76	0.522	<u>0.401</u>
Ours		22.34	0.732	0.286

Additionally, the per-scene optimization experiments are structured into two segments. In the initial segment, our objective is to assess the model’s performance post a finite series of fine-tuning iterations for each distinct scene. Specifically, we conducted two experimental sets, entailing 10k and 30k finetuning iterations per scenario. The pertinent outcomes and analyses are depicted in **Table 3**, illustrating that the model exhibits considerable performance enhancement following merely 10k iterations, indicative of its swift convergence properties. In the subsequent segment, We compared our model with previous models that failed to generalize under sparse input conditions. These methods rely on providing sparse input views for each scene and synthesizing

novel views through a limited number of iterations, without the ability to generalize across scenes. We compare these models by setting the same number of input views and fine-tuning 1k times. Verified through extensive experiments, our model is able to achieve higher quality view rendering with limited inputs, as shown in **Table 4**.

4.2 Ablation Experiments

In this paper, a series of ablation experiments were conducted to analyze the effectiveness and robustness of each module under extremely sparse input conditions. Initially, we eliminated the Transformer from the feature extraction module to assess its impact. Subsequently, we chose to remove both the Transformer in the feature extraction module and the Class Token (clt) in the ViT module. After removal, we select the input feature $fea_{n,1}$ corresponding to the input view closest to the target view as the final fusion density token $\{fea'_{n,1}\}$. In the final ablation study, we selected Multi-Head Attention as our feature fusion module, with the feature extraction module adopting our proposed hybrid architecture of encoder-decoder and Transformer. The specific experimental metrics comparison is shown in **Table 5**.

Table 5. Ablation study of key components on Synthetic Datasets [13].

Ablation Study	PSNR\uparrow	SSIM\uparrow	LPIPS\downarrow
w/o Transformer	22.92	0.871	0.173
w/o Transformer + w/o clt	22.58	0.867	0.176
w/o ViT	22.76	0.874	0.157
Full model Ours	22.96	0.876	0.155

5 Conclusion

We contend that under sparse view conditions, bolstering the extraction of global features aids the network in better deciphering the scene’s spatial relationships, thereby diminishing artifacts and enhancing novel view synthesis. Technically, our methodology incorporates two pivotal components: Firstly, leveraging our devised hybrid architecture combining an encoder-decoder with a Transformer, we extract comprehensive local and global feature information from sparse input views. This dual feature extraction enables the model to capture intricate scene details and textures, while global features assist in comprehending spatial relationships. Secondly, we deploy a Vision Transformer (ViT) sub-architecture for feature fusion, facilitating the effective amalgamation of features from diverse input views, thus constructing an integrated scene representation within the feature space. Based on this representation, our framework further predicts the density information of spatial points and employs the original NeRF’s volumetric rendering technique to synthesize the final novel view. Although our method significantly enhances the rendering quality, ViT has a huge parameter system, which leads to a decrease in inference speed and high computational complexity. In the future, we will try to use convolution with fewer parameters to replace ViT to achieve multi-view feature fusion.

References

1. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)
2. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12892–12901 (2022)
3. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
4. Kulháněk, J., Derner, E., Sattler, T., Babuška, R.: Viewformer: Nerf-free neural rendering from few images using transformers. In: European Conference on Computer Vision. pp. 198–216. Springer (2022)
5. Venkat, N., Agarwal, M., Singh, M., Tulsiani, S.: Geometry-biased transformers for novel view synthesis. arXiv Preprint arXiv:2301.04650 (2023)
6. Wynn, J., Turmukhambetov, D.: Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4180–4189 (2023)
7. Somraj, N., Soundararajan, R.: Vip-nerf: Visibility prior for sparse input neural radiance fields. arXiv Preprint arXiv:2305.00041 (2023)
8. Yang, J., Pavone, M., Wang, Y.: Freenerf: Improving few-shot neural rendering with free frequency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8254–8263 (2023)
9. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021)
10. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: Pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
11. Miralles, F., Posern, G., Zaromytidou, A.I., Treisman, R.: Actin dynamics control srf activity by regulation of its coactivator mal. *Cell* 113(3), 329–342 (2003)
12. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021)
13. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* 65(1), 99–106 (2021)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv Preprint arXiv:2010.11929 (2020)
15. Kirkland, E.J., Kirkland, E.J.: Bilinear interpolation. *Advanced Computing in Electron Microscopy* pp. 261–263 (2010)
16. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022)

17. Flynn, J., Broxton, M., Debevec, P., DuVall, M., Fyffe, G., Overbeck, R., Snavely, N., Tucker, R.: Deepview: View synthesis with learned gradient descent. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2367–2376 (2019)
18. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)* 38(4), 1–14 (2019)
19. Johari, M.M., Lepoittevin, Y., Fleuret, F.: Geonerf: Generalizing nerf with geometry priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18365–18375 (2022)
20. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5885–5894 (2021)
21. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. *Electronics Letters* 44(13), 800–801 (2008)
22. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4), 600–612 (2004)
23. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)