PVSeRF: Joint Pixel-, Voxel- and Surface-Aligned Radiance Field for Single-Image Novel View Synthesis

Xianggang Yu
FNii, CUHK-Shenzhen
xianggangyu@link.cuhk.edu.cn

Jiapeng Tang
Technische Universität München
jiapeng.tang@tum.de

Yipeng Qin
Cardiff University
qiny16@cardiff.ac.uk

Chenghong Li
FNii, CUHK-Shenzhen
chenghongli@link.cuhk.edu.cn

Xiaoguang Han*
SRIBD and FNii, CUHK-Shenzhen
hanxiaoguang@cuhk.edu.cn

Linchoao Bao
Tencent AI Lab
linchaobao@gmail.com

Shuguang Cui
SRIBD and FNii, CUHK-Shenzhen
shuguangcui@cuhk.edu.cn

ABSTRACT

We present PVSeRF, a learning framework that reconstructs neural radiance fields from single-view RGB images, for novel view synthesis. Previous solutions, such as pixelNeRF [68], rely only on pixel-aligned features and suffer from feature ambiguity issues. As a result, they struggle with the disentanglement of geometry and appearance, leading to implausible geometries and blurry results. To address this challenge, we propose to incorporate explicit geometry reasoning and combine it with pixel-aligned features for radiance field prediction. Specifically, in addition to pixel-aligned features, we further constrain the radiance field learning to be conditioned on i) voxel-aligned features learned from a coarse volumetric grid and ii) fine surface-aligned features extracted from a regressed point cloud. We show that the introduction of such geometry-aware features helps to achieve a better disentanglement between appearance and geometry, i.e. recovering more accurate geometries and synthesizing higher quality images of novel views. Extensive experiments against state-of-the-art methods on ShapeNet benchmarks demonstrate the superiority of our approach for single-image novel view synthesis.

KEYWORDS

Novel view synthesis, Single image, Neural radiance field

1 INTRODUCTION

Novel view synthesis is a long-standing problem in computer vision and graphics, which plays a crucial role in various practical applications, including gaming, movie production, and virtual/augment reality. Recently, it has made great strides thanks to the advances in differentiable neural rendering [37, 66], especially the neural radiance fields (NeRF) [31] that simplifies novel view synthesis to an optimization problem over a dense set of ground truth views. Although achieving impressive results, the vanilla NeRF suffers from several limitations: i) the dense views it strictly requires are not always available; ii) it is slow in inference due to the long optimization process; iii) each NeRF is dedicated to a specific scene and cannot be generalized to new ones.
To address these issues, follow-up works such as pixelNeRF [68], IBRNet [59], and GRF [56], proposed to predict neural radiance fields in a feed-forward manner. Taking pixelNeRF as an example, it tackles the shortcomings of NeRF by extending its network to be conditioned on scene priors learnt by a convolutional image encoder. These scene priors are represented by spatial feature maps that allow the mapping from a pair of query spatial point and viewing direction to their corresponding pixel-aligned features. In pixelNeRF, such a mapping is implemented by standard camera projection and bilinear interpolation. During inference, the scene priors are obtained via a forward pass through the image encoder and thus allow fast novel view synthesis from a single input view of diverse scenes. Although effective, pixelNeRF suffers from feature ambiguity issues that originate from the many-to-one mapping between queries and their corresponding pixel-aligned features. In other words, pixelNeRF naively assigns the same pixel-aligned features to different points in some novel view as long as these points overlap with each other in the input view, which can cause confusion (Fig. 2).

To clarify such ambiguity issues, we propose to incorporate explicit geometry reasoning and combine it with pixel-aligned features for radiance field prediction. Specifically, we leverage the recent success in single-view 3D reconstruction [4, 7, 10, 13, 30, 38, 49, 50] and inject rich geometry information into radiance field prediction by incorporating geometry-aware features of two shape representations: i) voxel-aligned features learned from a coarse volumetric grid and ii) fine surface-aligned features extracted from a regressed point cloud. Intuitively, such geometry-aware features augment pixel-aligned features with additional “dimensions”, thereby allowing previously ambiguous points to be separable. Furthermore, by constraining the radiance field learning on these geometry-aware features, our method not just synthesizes higher quality images of novel views, but also recover more accurate underlying geometries in radiance field, as witnessed in Fig. 1.

Our main contributions include:

- We propose a novel approach of learning neural radiance fields from single-view images jointly conditioned on pixel-, voxel-, surface-aligned features.
- We design an efficient way to alleviate the feature ambiguity issue of solely pixel-aligned features by incorporating explicit geometry reasoning via single-view 3D reconstruction.
- We propose a hybrid use of geometric features, including complementary coarse volumetric features and fine surface features.

2 RELATED WORK
2.1 Novel view synthesis and Neural radiance field

The task of novel view synthesis aims to generate new views of a scene from single or a set of sparse views. There are various kinds of approaches dedicated to this problem. Traditional methods [8, 12, 25] choose to estimate light fields and then render novel views. Recent years, with the advance of deep neural networks (DNN), a plethora of models are designed to learn novel view synthesis in an end-to-end manner. Pioneering methods [39, 53, 70] consider it as a image-to-image transformation problem and directly utilize 2D CNN to output novel views. These methods always cannot generate satisfactory results for viewpoints that are largely deviated from the given view.

Later work explore 3D-aware image synthesis and solve the inverse rendering problem via neural networks [11, 23, 35, 36, 47, 61, 71]. The common characteristic of this line of literature is that they recover the explicit or implicit 3D geometry and appearance properties first, then render novel views at desired camera viewpoints by means of differentiable rendering techniques [35] or generative models. Among these works, various 3D representations are employed. DeepVoxels [47] represents 3D scene properties by low-resolution volumetric feature grid lifted from 2D feature maps. Wiles et al. [61] use 3D surface features that are learned from the point cloud un-projected from the estimated depth map of the input view. Other approaches [11, 23, 36, 71] learn implicit 3D embedding that can be used to generate novel views of the same scene using unsupervised learning techniques.

Recently, witnessing the great success of neural radiance field (NeRF) [31], there has been an explosion of NeRF-based approaches for novel view synthesis[3, 6, 26–28, 33, 41, 45, 55, 56, 59, 67, 68]. There are two divisions in the prevalence of NeRF: 1) the first track tries to train scene-specific model for generating novel views of the scene [26–28, 33, 41, 45, 55, 67]. Specifically, they capture many diverse viewpoints of a scene, and optimizing a neural radiance field for that scene. Despite synthesizing high-fidelity novel views, these methods require longstanding optimization process and cannot generalize to new scenes. 2) the second track attempts to learn generalize neural radiance field across multiple scenes [3, 6, 56, 59, 68]. Among this, pixelNeRF [68] is the most relevant method to ours, which learns the scene priors conditioned on the pixel-aligned features, and can switch to new scenes flexibly. Although other methods [3, 6, 56, 59] can also be applied to novel scene through a
Figure 3: Overview of our PVSeRF framework. Given a single input image, we first 1) extract the spatial feature map using a fully convolutional image encoder $E$, 2) learn a volumetric grid through volume generator $G_V$, and 3) regress a surface point set of the object through a point set generator $G_S$. From volumetric grid and surface point set, we can learn voxel features and point-wise features respectively. Next, the 3D location, view direction and all corresponding features are directed into a MLP to predict density $\sigma$ and radiance $r$. Lastly, the volume rendering is used to accumulate the radiance prediction of points on the same ray to compute the final color values.
\[ T(t) = \exp \left( - \int_{t_n}^{t} \sigma(s) ds \right) . \] (2)

In practice, these integrals are approximated using the numerical quadrature rule [29]:
\[ \hat{C}(r) = \sum_{k=1}^{K} T_k (1 - \exp(-\sigma_k (t_{k+1} - t_k))) c_k, \]
with
\[ T_k = \exp \left( - \sum_{k' < k} \sigma_{k'} (t'_{k+1} - t'_{k}) \right). \] (3)

During training, the weights of a NeRF network are randomly initialized and optimized for an individual scene using a collection of RGB images, by minimizing a sole photometric loss \( L_{\text{photo}} = \sum_{r \in R} \| \hat{C}(r) - C(r) \|_2^2 \), in which \( r \in R \) is a set of randomly sampled rays from some images and \( C(r) \) is the ground truth color value of the pixel corresponding to ray \( r \).

### 3.2 Overview

Different from NeRF’s approach which must be optimized per-scene individually, our PVSeRF framework leverage the prior knowledge across multiple scenes and can reconstruct a neural radiance field from as little as a single image which is similar to pixelNeRF [68]. Specifically, given a single calibrated image \( I \) with its corresponding intrinsic \( K \) and extrinsic parameters (rotation \( R \) and translation \( t \)), our PVSeRF aims to learn a neural network for radiance field reconstruction:
\[ \sigma, c = \text{PVSeRF}(X, d; I, [Rt], K) \] (4)

where \( X \in \mathbb{R}^3 \) represents the 3D location, \( d \in \mathbb{R}^2 \) is the viewing direction, \( \sigma \) is the volume density at \( X \), and \( c \) is the predicted color at \( X \) depending on the viewing direction \( d \). By accumulating the \( \sigma \) and \( c \) of multiple points sampled on the ray defined by \( X \) and \( d \), we can obtain the color values of all pixels in a target view image \( I_t \) via differentiable rendering, thereby enabling novel view synthesis.

The distinct advantage of our PVSeRF is that it addresses the feature ambiguity issue of pixelNeRF [68] by a novel geometric regularization using both voxel- and surface-aligned features. As aforementioned, pixelNeRF’s feature ambiguity issue stems from the fact that its network is solely conditioned on the 2D pixel-aligned features where multiple query 3D points are mapped to a single location. To clarify this ambiguity, we propose to augment the 2D pixel-aligned features with complementary 3D geometric features for radiance field construction. As Fig. 3 shows, in addition to the pixel-aligned features, our method incorporates i) voxel-aligned and ii) surface-aligned features into radiance field prediction. Specifically,

- We follow [68] and extract the pixel-aligned features \( f_i \) of a query point \( X \) by projecting it with \([Rt]\) and \( K \) to the 2D image coordinates \( x \), indexing the multi-scale feature maps of an input image \( I \) extracted by a fully-convolutional image encoder \( E \).
- We extract the voxel-aligned features \( f_v \) of a query point \( X \) by trilinearly interpolating \( X \) in a low-resolution volumetric feature \( F_v \) learnt from the input image \( I \) using a volume generator \( G_v \). Note that \( f_v \) only captures coarse geometry contexts of the scene due to the low-resolution nature of \( F_v \).
- To capture the geometric information on surface, we extract the fine-grained surface-aligned features \( f_s \) of a query point \( X \) as the weighted sum of the associated features \( f_y \) of its \( K \) nearest neighbors in a point cloud \( S \), which is reconstructed from the input image \( I \) using a point set generator \( G_s \).

Thus, our PVSeRF is conditioned on \( f_i, f_v, \) and \( f_s \) and can be reformulated as:
\[ \sigma, c = \text{PVSeRF}(X, d; f_i \oplus f_v \oplus f_s) \] (5)

where \( \oplus \) denotes a concatenation operation. Thanks to the incorporation of \( f_v \) and \( f_s \), the previously ambiguous points that share the same \( f_i \) are now separable by the concatenation \( f_i \oplus f_v \oplus f_s \). We present more details about each component of our method as follows.

### 3.3 Feature Extraction

**Pixel-aligned Features** Following pixelNeRF [68], we also use pixel-aligned features that contain fine-grained details about the scene’s geometry and appearance properties to learn neural radiance fields. Given an input image \( I \in \mathbb{R}^{H \times W \times 3} \), we employ a fully-convolutional image encoder \( E \) implemented by ResNet-34 [16] to extract its multi-scale feature maps \( \{F^0_1, F^1_1, F^2_1, F^3_1\} \), which are the intermediate features at ‘conv1’, ‘layer1’, ‘layer2’, and ‘layer3’ of ResNet-34 but upsampled to the size of the input image \( I \). Then, we acquire the pixel-aligned feature vector \( f_i \) of a query 3D point \( X \) by projecting \( X \) to the 2D image coordinates \( x \), and bilinearly interpolating the feature maps concatenated by \( \{F^0_1, F^1_1, F^2_1, F^3_1\} \) through \( B \):
\[ f_i = B(F^0_1 \oplus F^1_1 \oplus F^2_1 \oplus F^3_1, K[Rt]X) \] (6)

where \( \oplus \) represents feature concatenation. However, \( K[Rt] \) may project multiple 3D points \( X \) along the viewing direction of input image to a single position on the 2D image coordinates, leading to ambiguous \( f_i \) and blurry synthesized novel views. To clarify such ambiguity, we propose to augment \( f_i \) with complementary geometric features, including both coarse voxel-aligned features learned from a volumetric grid, and fine surface-aligned features extracted from a regressed point cloud.

**Voxel-aligned Features** We compute the voxel-aligned feature \( f_v \) with respect to \( X \) as follows. First, we reconstruct a volumetric feature grid \( F_v \in \mathbb{R}^{32 \times 32 \times 32} \) from the input image \( I \) using a volume generator consisting of a VGG-16 [46] image encoder and a 3D CNN decoder. Then, we have:
\[ f_v = T(F_v, \Omega(X)) \] (7)

where \( T \) is a multi-scale trilinear interpolation inspired by GeoPiFu [17] and IFNet [5]. \( \Omega(X) \) is a point set around \( X \):
\[ \Omega(X) = \{ X + s \cdot n | n = (1, 0, 0), (0, 1, 0), (0, 0, 1), \ldots \} \] (8)

where \( s \in \mathbb{R} \) is the step length, \( n \in \mathbb{R}^3 \) represents the unit vectors defined along the three axes in a Cartesian coordinate system. Intuitively, \( f_v \) is a concatenation of all queried feature vectors at points in \( \Omega(X) \) that are trilinearly interpolated from \( F_v \).

**Surface-aligned Features** Although they capture a global context about the shape of a 3D object, voxel-aligned features are queried from a low-resolution volumetric grid and thus lack geometric information on surface. As a complement, we introduce surface-aligned features that capture fine details of surface to facilitate radiance field
We parameterize our PVSeRF framework using a MLP \( f \) which regresses the volume density \( \sigma \) and view-dependent radiance \( r \) from the 3D coordinates of a query point \( X \), a viewing direction \( d \), and the corresponding pixel-, voxel-, and surface-aligned features (i.e. \( f_{\uparrow} \), \( f_{\downarrow} \), and \( f_{\circ} \)) extracted from the input single-view image \( I \):

\[
\sigma, r = f(\gamma_{m}(X), \gamma_{n}(d); f_{\uparrow} \odot f_{\downarrow} \odot f_{\circ})
\]  

where \( \gamma_{m} \) and \( \gamma_{n} \) are position encoding functions \([31, 58]\) applied to \( X \) and \( d \) respectively, which alleviates the positional bias inherent in Cartesian coordinates without sacrificing their discrepancy in-between. Specifically, \( \gamma \) maps Cartesian coordinates from \( \mathbb{R} \) into a high dimensional space \( \mathbb{R}^{2L} \):

\[
y_{L}(p) = (\sin(2^{0} \pi p), \cos(2^{0} \pi p), \ldots, \sin(2^{L-1} \pi p), \cos(2^{L-1} \pi p))
\]  

where \( \gamma(\cdot) \) is applied separately to each component of vector \( p \). With the constructed radiance field represented by \( \sigma \) and \( r \), we render novel view images via the numerical quadrature approximation of differentiable volume rendering techniques which illustrated in Section 3.1.

### 3.5 Loss Functions

Corresponding to our pixel-, voxel- and surface-aligned features, we train our model using three different loss functions as follows.

**RGB Rendering Loss** Similar to existing works in the NeRF series, we use \( L2 \) rendering loss as the main loss function. It constrains that the rendered color value of each ray should be consistent with the corresponding ground-truth pixel value. Thus, we have:

\[
L_{r} = \|\hat{C}(r) - C(r)\|_{2}^{2}
\]
Volume Reconstruction Loss
To learn volumetric features $F_v$, we add a 3D convolutional layer after $F_v$ to estimate a low-resolution occupancy volume $V \in \mathbb{R}^{32 \times 32 \times 32}$, whose ground-truth label is $V^*$. Then, we apply a standard binary cross-entropy loss and have:

$$L_v = \sum_{i \in [32]^3} V'(i) \log V(i) + (1 - V'(i)) \log(1 - V(i))$$  \hspace{1cm} (14)

Point Regression Loss
We employ the Chamfer distance to constraint our point set generation and have:

$$L_p = \sum_{q \in S} \min_{q' \in S^*} ||q - q'||^2 + \sum_{q' \in S^*} \min_{q \in S} ||q - q'||^2$$  \hspace{1cm} (15)

where $S$ is the predicted point set and $S^*$ is its corresponding ground truth.

Overall Loss Function
Our overall loss function is:

$$L = \lambda_1 * L_v + \lambda_2 * L_v + \lambda_3 * L_p$$

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are weighting parameters.

4 EXPERIMENTS

To demonstrate the superiority of our PVSeRF, we first compare it against state-of-the-art methods on two single-image novel view synthesis tasks, i.e. category-agnostic view synthesis and category-specific view synthesis. Then, we evaluate our approach on real images, using the collected real-world cars images from [24]. To provide supervision for volume reconstruction and point set regression, we convert each ground-truth mesh to a point set of size 2048 and a volumetric grid of resolution 32.

Implementation Details
We implement our model with PyTorch [42]. Details of the network architecture are presented in the supplementary material. The training process of our approach consists of two stages: i) we pre-train the volume generator $G_V$ and the point set generator $G_S$ respectively using loss functions defined in Eq. 14 and Eq. 15. Specifically, $G_V$ is trained with an initial learning rate of $10^{-3}$ and a batch size of 64 for 250 epochs. The learning rate drops by a factor of 5 after 150 epochs. $G_S$ is trained with an initial learning rate of $10^{-5}$ and a batch size of 4 for 10 epochs. The learning rate drops by a factor of 3 after 5 and 8 epochs. ii) we fine-tune the whole network for 400 epochs. We set the learning rate as $10^{-4}$ and the batch size as 4. We use an Adam [21] optimizer for all the training mentioned above. We empirically set the multi-scale trilinear interpolation step length $s = 0.0722$, the number of nearest neighbors $K = 5$, the number of frequencies of positional encoding for $X, d$ as $m = 6, n = 0$, and the weights for loss function as $\lambda_1 = \lambda_2 = \lambda_3 \approx 1$.

Evaluation Protocol
Following the community standards [31, 37, 48], we use peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [60] to measure the quality of the synthesized novel views. We also use LPIPS [69] that has been shown to be closer to human perception.

4.1 Category-agnostic View Synthesis

Category-agnostic novel view synthesis aims to learn object priors that can generalize across multiple categories.

Baselines
We compare our method against three closely-related state-of-the-art methods: SRN [48], DVR [37] and pixelNeRF [68], which are applicable to synthesize novel views for all categories. For DVR and pixelNeRF, we use pretrained models from their official Github repositories. For SRN [48], we use the model trained by [68] to make it comparable with [48, 68]. All methods are trained

### Table 2: Quantitative comparison on category-specific view synthesis.

| Method          | PSNR↑ | SSIM↑ |
|-----------------|-------|-------|
| TCO [52]        | 21.27 | 0.88  |
| dGQN [9]        | 21.59 | 0.87  |
| Chairs SRN [48] | 22.89 | 0.89  |
| pixelNeRF [68]  | 23.72 | 0.91  |
| Ours            | 23.33 | 0.91  |
| Cars SRN [48]   | 22.25 | 0.89  |
| pixelNeRF [68]  | 23.17 | 0.90  |
| Ours            | 22.98 | 0.90  |

1Niemeyer et al. [37]: https://github.com/autonomousvision/differential_volumetric_rendering, Yu et al. [68]: https://github.com/sxyu/pixel-nerf.
Figure 6: Qualitative comparison on category-specific view synthesis. The performance of our method is comparable to that of the state-of-the-art pixNeRF [68].

using the same dataset and settings introduced in Sec. 4. To facilitate a fair comparison, we follow the random view indices provided by pixNeRF and select the input view for each test object accordingly.

Results As Fig. 4 shows, our method outperforms all previous methods by synthesizing more detailed novel views. In addition, it can be observed that i) the two baseline methods, DVR [37] and SRN [48], tend to generate blurry images and distorted geometries; ii) pixNeRF [68] shows blurry and inconsistent appearance. The quantitative results in Table 1 further justify the superiority of our method against all baselines in terms of the mean values of PSNR, SSIM and LPIPS metrics. Notably, the PSNR of our approach attains a significant improvement over the second best method by 0.68.

4.2 Category-specific View Synthesis

For category-specific view synthesis, all methods are trained on the chair or car categories of ShapeNet [2].

Baselines We choose SRN [47] and pixNeRF [68] as the baseline methods. We also report the quantitative results from TCO [52] and dGQN [9] provided by [48], to keep in line with prior arts.

Results We show the quantitative and qualitative results in Table 2 and Fig. 6 respectively. It can be observed that the performance of our method is comparable to the state-of-the-art method [68] both qualitatively and quantitatively. Such comparable results indicate that the advantages of our method are not significant in some neural rendering cases. We carefully investigate the results and ascribe this to the invalidity of explicit geometry reasoning in some cases (Fig. 5). Since the renderings provided by [48] contain many challenging camera viewpoints, the explicit geometry reasoning from single-view becomes a more challenging problem. We postpone the discussion of this phenomenon to Sec. 5.

4.3 Novel View Synthesis on Real Images

To highlight the generalization ability of our method, we evaluate our pretrained models directly on real images without any finetuning. Specifically, we first take the images from the Stanford cars dataset [24] and apply the PointRend model [22] to mask their clutter backgrounds. Then, we feed the preprocessed images into a category-specific model of ShapeNet “cars” to predict novel views. As Fig. 7 shows, our method can not only synthesize visually compelling novel views, but also infer accurate geometries. This effectively demonstrates the excellent generalization performance of our method on real image as it is only trained on synthetic images.

4.4 Ablation Study

To validate the effectiveness of each proposed component, we conduct an ablation study on our method, yielding three variants: i) w/o surface-aligned feature, in which only pixel- and voxel-aligned features are incorporated; ii) w/o voxel-aligned feature, where the radiance field is conditioned only on pixel- and surface-aligned features; iii) w/o joint training, in which we fix all feature extractors and solely train the radiance field predictor $f$. As Table 3 shows, it can be observed that the w/o joint training variant constantly performs the worst among all variants. This demonstrates that the joint learning of pixel-, voxel- and surface-aligned features is crucial in our explicit geometric reasoning. In addition, the performance of the w/o surface-aligned variant is always worse than the w/o voxel-aligned feature variant, as the voxel-aligned features queried from a low-resolution volume are better at capturing global geometry contexts. Our full method achieves the best performance among all variants, which validates the effectiveness of employing a hybrid of geometric features that complement each other. This is also demonstrated in Fig. 8.

5 CONCLUSION

For the task of novel view synthesis from single-view RGB images, we present PVSeRF, a novel learning framework that reconstructs neural radiance fields conditioned on joint pixel-, voxel-, and surface-aligned features. By augmenting hybrid geometric features with image features, we effectively address the feature confusion issue of pixel-aligned features. Compare to previous arts, our framework

2We use a PointNet++ [44] model trained on PartNet [32] segmentation task as a point-feature extractor.
Figure 7: Novel view synthesis results on real car images. Although extensively trained on synthetic data, our method can easily generalize to real single-view images, and produce plausible view synthesis results and underlying geometries.

|                | plane | bench | cbnt. | car | chair | disp. | lamp | spkr. | rifle | sofa | table | phone | boat | mean |
|----------------|-------|-------|-------|-----|-------|-------|------|-------|-------|------|-------|-------|------|------|
| w/o joint      | 29.03 | 25.18 | 26.12 | 25.82 | 21.97 | 22.25 | 26.33 | 22.19 | 25.18 | 24.27 | 24.67 | 27.54 | 25.16 |
| w/o surface-aligned | 30.82 | 27.00 | 28.31 | 27.67 | 24.05 | 24.33 | 28.73 | 24.33 | 30.63 | 26.97 | 26.27 | 26.85 | 29.58 | 27.15 |
| ↑ PSNR         | 30.83 | 27.14 | 28.40 | 27.93 | 24.35 | 24.66 | 29.10 | 24.75 | 31.05 | 27.29 | 26.48 | 27.01 | 29.61 | 27.38 |
| w/o voxel-aligned | 31.32 | 27.43 | 28.40 | 28.12 | 24.37 | 24.61 | 28.73 | 24.44 | 30.82 | 27.42 | 26.60 | 27.42 | 29.92 | 27.48 |
| Ours           | 30.83 | 27.14 | 28.40 | 27.93 | 24.35 | 24.66 | 29.10 | 24.75 | 31.05 | 27.29 | 26.48 | 27.01 | 29.61 | 27.38 |

Table 3: Quantitative comparison of ablation studies. Our joint method that employs complementary coarse volumetric features and fine surface features achieves the best performance. Whereas, removing any part of the proposed method will cause more or less deterioration.

Figure 8: Illustration of the complementary properties of point set and volumes. We randomly show several predicted geometries. It can be seen that these two representations exhibit reciprocal behaviors: the missing parts of volumetric grid are spanned by point set, while the regions where the point set is too sparse are occupied by volumes.

The work was supported in part by the National Key R&D Program of China with grant No. 2018YFB1800800, the Basic Research Project No. HZQB-KCZYRZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, and by the Guangdong Provinical Key Laboratory of Future Networks of Intelligence (Grant No. 2022B121200192). It was also supported by the National Key R&D Program of China with grant No. 2019YFE0110100 and Shenzhen General Project (No. JCYJ20190814112007258).
[49] Jiapeng Tang, Xiaoguang Han, Junyi Pan, Kui Jia, and Xin Tong. 2019. A skeleton-bridged deep learning approach for generating meshes of complex topologies from single rgb images. In CVPR.

[50] Jiapeng Tang, Xiaoguang Han, Mingkui Tan, Xin Tong, and Kui Jia. 2020. SkeletonNet: A Topology-Preserving Solution for Learning Mesh Reconstruction of Object Surfaces from RGB Images. arXiv preprint arXiv:2008.05742 (2020).

[51] Jiapeng Tang, Jiabo Lei, Dan Xu, Feiyong Ma, Kui Jia, and Lei Zhang. 2021. SA-ConvONet: Sign-Agnostic Optimization of Convolutional Occupancy Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 6504–6513.

[52] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. 2015. Single-view to Multi-view: Reconstructing Unseen Views with a Convolutional Network. ArXiv abs/1511.06702 (2015).

[53] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. 2016. Multi-view 3d models from single images with a convolutional network. In European Conference on Computer Vision. Springer, 322–337.

[54] Engin Tola, Christoph Strecha, and Pascal Fua. 2012. Efficient large-scale multi-view stereo for ultra high-resolution image sets. Machine Vision and Applications 23, 5 (2012), 903–920.

[55] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. 2020. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Deforming Scene from Monocular Video. https://arxiv.org/abs/2012.12247 (2020).

[56] Alex Treivthick and Bo Yang. 2020. Grf: Learning a general radiance field for 3d scene representation and rendering. arXiv preprint arXiv:2010.04595 (2020).

[57] Richard Tucker and Noah Snavely. 2020. Single-view view synthesis with multi-plane images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 551–560.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems. 5998–6008.

[59] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brulla, Noah Snavely, and Thomas Funkhouser. 2021. Brnet: Learning multi-view image-based rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4690–4699.

[60] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 4 (2004), 600–612.

[61] Olivia Wiles, Georgia Okioxiari, Richard Szeliski, and Justin Johnson. 2020. Synski: End-to-end view synthesis from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7467–7477.

[62] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. 2021. Nex: Real-time view synthesis with neural basis expansion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8534–8543.

[63] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In NeurIPS.

[64] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1912–1920.

[65] Yaqing Yang, Chen Feng, Yuru Shen, and Dong Tian. 2018. Foldingnet: Point cloud auto-encoder via deep grid deformation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 206–215.

[66] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. arXiv preprint arXiv:2003.09832 (2020).

[67] Alex Yu, Rui Long Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021. Plenoctrees for real-time rendering of neural radiance fields. arXiv preprint arXiv:2103.14024 (2021).

[68] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4578–4587.

[69] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition. 586–595.

[70] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. 2016. View synthesis by appearance flow. In European conference on computer vision. Springer, 286–301.

[71] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Joshua B Tenenbaum, and Bill Freeman. 2018. Visual Object Networks: Image Generation with Disentangled 3D Representations. In NeurIPS.