Painting 3D Nature in 2D: View Synthesis of Natural Scenes from a Single Semantic Mask

Shangzhan Zhang[1*] Sida Peng[1*] Tianrun Chen[1] Linzhan Mou[1] Haotong Lin[1*]
Kaicheng Yu[2] Yiyi Liao[1] Xiaowei Zhou[1*†]
[1]Zhejiang University [2]Alibaba Group

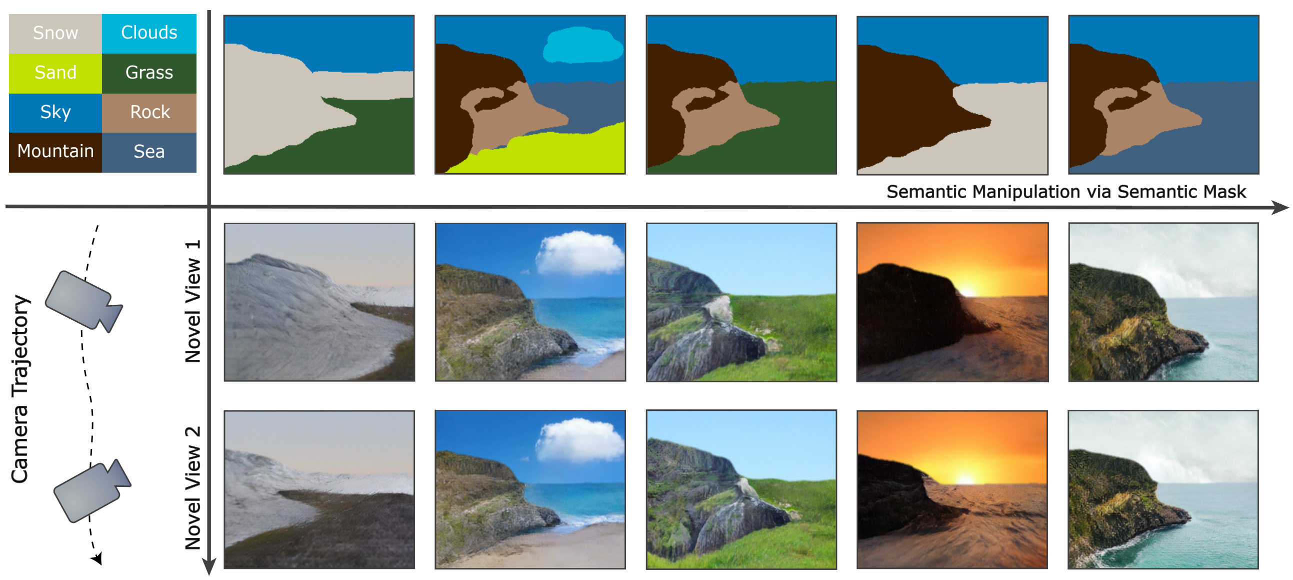

Figure 1. Given only a single semantic map as input (the first row), our approach is able to generate neural fields for view synthesis of natural scenes. Photorealistic images can be rendered via neural fields (the last two rows).

Abstract

We introduce a novel approach that takes a single semantic mask as input to synthesize multi-view consistent color images of natural scenes, trained with a collection of single images from the Internet. Prior works on 3D-aware image synthesis either require multi-view supervision or learning category-level prior for specific classes of objects, which are inapplicable to natural scenes. Our key idea to solve this challenge is to use a semantic field as the intermediate representation, which is easier to reconstruct from an input semantic mask and then translated to a radiance field with the assistance of off-the-shelf semantic image synthesis models. Experiments show that our method outperforms baseline methods and produces photorealistic and multi-view consistent videos of a variety of natural scenes. The project website is https://zju3dv.github.io/paintingnature/.

*Affiliated with the State Key Lab of CAD&CG, Zhejiang University.
†Corresponding author: Xiaowei Zhou.

1. Introduction

Natural scenes are indispensable content in many applications such as film production and video games. This work focuses on a specific setting of synthesizing novel views of natural scenes given a single semantic mask, which enables us to generate 3D contents by editing 2D semantic masks. With the development of deep generative models, 2D semantic image synthesis methods [24, 46, 61, 66] have achieved impressive advances. However, they do not consider the underlying 3D structure and cannot generate multi-view consistent free-viewpoint videos.

To address this problem, a straightforward approach is first utilizing semantics-driven image generator like SPADE [46] to synthesize an image from the input semantic mask and then predicting novel views based on the generated image. Although the existing single-view view synthesis methods [31, 34, 45, 52, 67, 70] achieve impressive rendering results, they typically require training networks on posed multi-view images. Compared to urban or indoor scenes, learning to synthesize natural scenes is a challenging task as it is difficult to collect 3D data or posed videos of natural scenes for training, as demonstrated in [32], making the aforementioned methods not applicable. AdaMPI [17] designs a training strategy to learn the view synthesis network on single-view image collections. It warps images to random novel views and warps them back to the original view. An inpainting network is trained to fill the holes in disocclusion regions to match the original images. After training, the inpainting network is used to generate pseudo multi-view images for training a view synthesis network. Our experimental results in Section 4.5 show that the inpainting network struggles to output high-quality image contents in missing regions under large viewpoint changes, thus limiting the rendering quality.

In this paper, we propose a novel framework for semantics-guided view synthesis of natural scenes by learning prior from single-view image collections. Based on the observation that semantic masks have much lower complexity than images, we divide this task into two simpler subproblems: we first generate semantic masks at novel views and then translate them to RGB images through SPADE. For view synthesis of semantic masks, the input semantic mask is first translated to a color image by SPADE, and a depth map is predicted from the color image by a depth estimator [50]. Then, the input semantic mask is warped to novel views using the predicted depth map and refined by an inpainting network trained by a self-supervised learning strategy on single-view image collections. Our experiments show that, in contrast to images, the novel view synthesis of semantic masks is much easier to learn by the network.

It is observed that semantic masks generated by the inpainting network tend to be view-inconsistent. As a result, SPADE could generate quite different contents in these regions even when the inconsistency is minor between the semantic masks. Fig. 4 presents two examples. To solve this issue, we learn a neural semantic field to fuse and denoise these semantic masks for better multi-view consistency. Finally, we translate the multi-view semantic masks to color images by SPADE and reconstruct a neural scene representation for view-consistent rendering.

Extensive experiments are conducted on the LHQ dataset [58], a widely-used benchmark dataset for semantic image synthesis. The results demonstrate that our approach significantly outperforms baseline methods both qualitatively and quantitatively. We also show that by editing the input semantic mask, our approach is capable of generating various high-quality rendering results of natural scenes, as shown in Fig. 1.

2. Related Work

Semantics-guided view synthesis. This task takes a single 2D semantic mask as input and outputs free-viewpoint videos of the 3D scene. Only a few works attempt to tackle this challenging task. GVS [16] and SVS [22] can generate MPI representations [76] from a single semantic mask and render free-viewpoint videos. However, they need to train their model on datasets of posed videos [4, 10, 76]. The problem with this training strategy is that obtaining a large number of videos with calibrated camera poses in real life could be expensive and limit the diversity of training data. Qiao et al. [49] mainly focus on novel-view scene layout generation and their model is also trained on posed images, while we concentrate on rendering consistent RGB images with single image collections for training.

Single-image view synthesis. Recently, many works have focused on single-view view synthesis with monocular RGB image as the input to generate free-viewpoint videos [21, 29, 34, 43, 51, 52, 56, 63, 67]. Among existing works, some use explicit 3D representations, such as layered depth image (LDI) [19, 55] and multi-plane image (MPI) [76], which can capture visible contents and infer disocclusion regions. Another line of research predicts neural radiance fields [36] from a single image. For example, PixelNeRF [70] extracts image features using 2D CNN and constructs an aligned feature field to render novel views. Li *et al*. [31] combine MPI and NeRF representations and learn a continuous 3D field. These works learn 3D representations to perform novel-view synthesis, and the learning is mainly based on multi-view images or posed videos for supervision. However, similar to the dataset constraint in semantic view synthesis, large-scale multi-view datasets are rare, thus bringing challenges for high-quality view synthesis for natural scenes.

Some methods have utilized single-view image collections to train a neural network that performs novel view synthesis given a single-view image. For example, [29] and [56] use a monocular depth estimation network to construct LDI representation and leverage the inpainting network to synthesize disocclusion content to perform view synthesis. AdaMPI [17] proposes a warp-back training strategy to train MPI on the COCO [5] datasets. Li *et al*. [32] propose a cycle-rendering strategy to train their networks on single image collections. Still, no prior works have attempted to train models solely using single-view image collections while performing novel view synthesis from a single semantic mask. A solution to leverage existing single-view image collections [58] to learn semantics-guided semantic view synthesis is extending the above approaches [17,56] through a two-stage scheme: first converting the semantic mask to an RGB image and then applying the view synthesis model. However, this solution does not fully utilize semantic information, thus leading to limited performance.

Neural radiance fields. Neural radiance fields [36] and its subsequent works significantly advance the realm of novel view synthesis [1, 2, 40, 68, 73] and 3D reconstruction [11,44,74]. While the above works focus on rendering realistic novel view images or reconstructing accurate 3D geometry, Semantic-NeRF [75] and later works [13,30,62] use NeRFs as a powerful 3D fusion tool to fuse 2D semantic information. Some works currently aim to exploit 2D pretrained models to learn a priori knowledge for the neural field. Examples include creating 3D objects driven by text [25,48,64], animating NeRFs by audio signals [15,35], and stylizing scenes [23,39,72]. In contrast to previous approaches, this paper exploits 2D pretrained models to generate neural fields based on a single semantic mask. Another line of work [9, 59, 60] extends 3D-aware generative modeling [7, 41, 42, 53, 54] to edit 3D appearance and geometry via a semantic mask, but they mainly conduct their experiments on object-centric datasets (e.g., FFHQ [28]) with known camera distribution and fail to generate complex scenes on the nature scene datasets. A closely related work GANcraft [18] and its extended work [26] use image-to-image translation techniques [24,46,61,66] to synthesize pseudo ground truths and discriminators to make generative free-viewpoint videos more realistic from 3D semantic labels. But they need 3D semantic labels to render consistent 2D semantic masks and are inapplicable when only a single semantic mask is available as input.

Image-to-image translations Image-to-image translations [24, 46, 61, 66] have made tremendous development and can synthesize realistic natural images. Pix2Pix [24] firstly leverages conditional GAN [37] to improve the performance of semantic image synthesis. The performance of image synthesis is significantly enhanced by SPADE [46], which proposes a spatial-varying normalization layer. OASIS [61] proposes a novel discriminator for semantic image synthesis [27]. However, these works focus on single-image synthesis, which are not able to generate multi-view images of 3D scenes.

3. Methods

Our goal is to perform photorealistic view synthesis of natural scenes, given a single semantic mask, by learning prior from single-view image collections. To this end, we divide this task into two simpler sub-problems. Section 3.1 first introduces how to generate multi-view consistent semantic masks from the given semantic mask. Then, Section 3.2 discusses how to translate the multi-view semantic masks to RGB images by SPADE and recover a neural scene representation for view-consistent rendering.

3.1. Generating view-consistent semantic masks

Fig. 2 illustrates the overview of generating multi-view consistent semantic masks from a single semantic mask. Specifically, we first warp the given semantic mask to novel views. Then, an inpainting network is utilized to fill in the disocclusion areas of the warped semantic mask at each novel view. After obtaining multiple infilled semantic masks at different viewpoints, we recover a neural semantic field that can fuse and denoise the multi-view semantic information. Finally, multi-view semantic masks can be obtained by the semantic field.

Warping semantic mask. Our approach warps the given semantic mask to novel views through the depth-based warping technique. We first convert the input semantic map to the corresponding RGB image using SPADE [46], and then use a monocular depth estimation network [50] to predict the depth map from the generated RGB image. Then, a 3D triangular mesh is constructed based on the predicted depth following Shih *et al*. [56]. The semantic mask is lifted to the mesh, whose vertices' color is assigned as the corresponding semantic label. The generated 3D triangular mesh in such a manner may contain spurious edges due to the depth discontinuities in the depth map. To solve this problem, we remove the edges whose vertices are far from each other. Eventually, we warp the semantic mask to the novel views using a mesh renderer following [56].

Semantic mask inpainting. Directly warping the given semantic mask to novel views brings many holes in disocclusion regions. To inpaint the missing contents, we train a semantic inpainting network on single-view natural image collections [58] using the self-supervised technique in [17]. Fig. 3 shows our training strategy for semantic inpainting networks. Specifically, we first use a pre-trained image segmentation model [8] and a monocular depth estimation model [50] to produce semantic masks and depth maps for the natural images. At each training iteration, an image is randomly sampled from the dataset as the source image $\mathbf{I}_i$. We then use the corresponding depth map $\hat{\mathbf{D}}_i$ to warp the original semantic mask $\mathbf{S}_i$ to a random target view j, producing warped semantic mask $\mathbf{S}_{i\to j}$ and depth map $\hat{\mathbf{D}}_{i\to j}$ at target view j. Next, we warp semantic mask $\mathbf{S}_{i\to j}$ back to the source view using depth map $\hat{\mathbf{D}}_{i\to j}$, which generates a semantic mask $\mathbf{S}_{i\to j\to i}$ with holes. Finally, we input $\mathbf{S}_{i\to j\to i}$ into the semantic inpainting network and train the network to infill these holes, which is supervised with the original semantic mask $\mathbf{S}_i$.

At test time, we first randomly sample a set of viewpoints, then warp the given semantic mask $\mathbf{S}_0$ to these view-

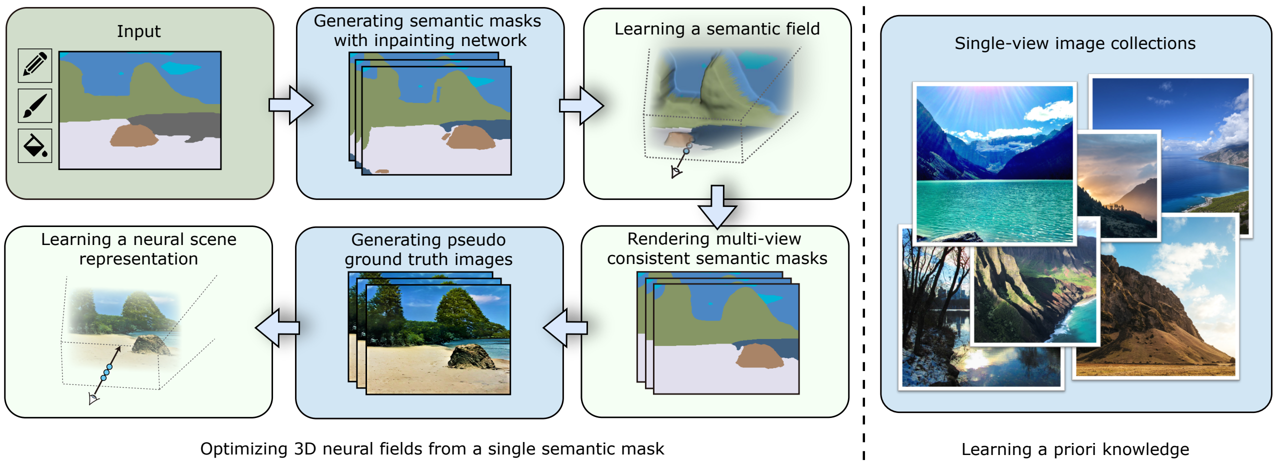

Figure 2. **Illustration of our pipeline. Left**: Our pipeline can be divided into two steps: we first generate multi-view semantic masks with an inpainting network and then convert semantic masks to RGB images using SPADE. In order to denoise and fuse semantic information, a semantic field is learned for rendering multi-view consistent masks. Finally, a neural scene representation is optimized to fuse appearance information provided by SPADE, which enables view-consistent rendering. **Right**: Our semantic inpainting network and SPADE are trained on single-view image collections.

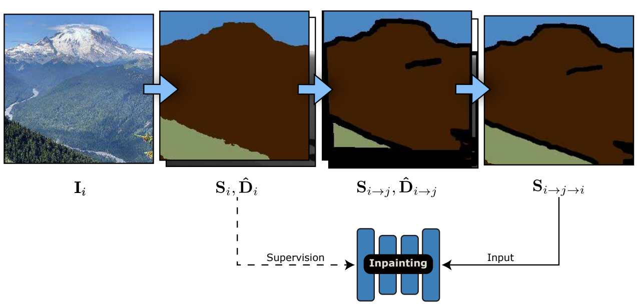

Figure 3. **Training a semantic inpainting network.** We produce a semantic mask $\mathbf{S}_i$ and a depth map $\hat{\mathbf{D}}_i$ from a randomly sampled image $\mathbf{I}_i$, then warp them to a random novel view generating $\mathbf{S}_{i\to j}$ and $\hat{\mathbf{D}}_{i\to j}$, and warp $\mathbf{S}_{i\to j}$ back to the source view to generate $\mathbf{S}_{i\to j\to i}$ with holes. Our semantic inpainting network takes the $\mathbf{S}_{i\to j\to i}$ as input and is trained to recover the $\mathbf{S}_i$.

points to generate warped semantic masks, and finally apply the inpainting network to fill in their disocclusion regions to generate the infilled multi-view semantic masks.

Semantic field fusion. We observed that the infilled semantic masks are not view-consistent. Although the artifact regions in semantic masks seem trivial, the generated images from SPADE could be very different on these regions across different viewpoints. Fig. 4 presents an example. To tackle this problem, a semantic field is introduced to fuse and denoise infilled semantic masks. We adopt a continuous neural field to represent the semantics and geometry of a 3D scene, similar to [14]. For any query point $\mathbf{x}$ in 3D space, an MLP network f_θ maps it to an SDF value d and an intermediate feature $\mathbf{z}$, and another MLP network f_ϕ maps $\mathbf{z}$ to a semantic logits $\mathbf{s}$. The neural field is defined as:

$$\begin{aligned} f_\theta &: \mathbf{x} \in \mathbb{R}^3 \mapsto (d \in \mathbb{R}, \mathbf{z} \in \mathbb{R}^c) \\ f_\phi &: \mathbf{z} \in \mathbb{R}^c \mapsto \mathbf{s} \in \mathbb{R}^{M_s}, \end{aligned} \tag{1}$$

where M_s denotes the number of semantic classes. We render the semantic field into semantic logits and depth through the SDF-based volume rendering [65, 69].

Considering that the sky is very distant from the foreground, we handle the foreground and the sky separately, following the practice in [18]. The sky is assumed to be a distant 2D plane, and the semantic probability of the sky is defined as a constant one-hot vector $\mathbf{P}_{\text{sky}}$. The final semantic probability is formulated as:

$$\mathbf{Y}(\mathbf{r}) = \mathbf{P}_{\text{fg}}(\mathbf{r}) T_{\text{fg}}(\mathbf{r}) + (1 - T_{\text{fg}}(\mathbf{r}))\mathbf{P}_{\text{sky}}, \tag{2}$$

where T_{fg} is the accumulated transmittance of the foreground along camera ray $\mathbf{r}$, and $\mathbf{P}_{\text{fg}}$ is the semantic probability obtained by applying the softmax layer to the rendered semantic logits.

To learn the semantic field, the cross entropy loss is applied to compare the rendered semantic probability $\mathbf{P}(\mathbf{r})$ and the semantic probability $\mathbf{P}^*(\mathbf{r})$ provided by the infilled semantic masks:

$$\mathcal{L}_{\mathbf{P}} = -\sum_{\mathbf{r}\in\mathcal{R}} \sum_{k=1}^{M_s} \mathbf{P}_k^*(\mathbf{r}) \log \mathbf{P}_k(\mathbf{r}). \tag{3}$$

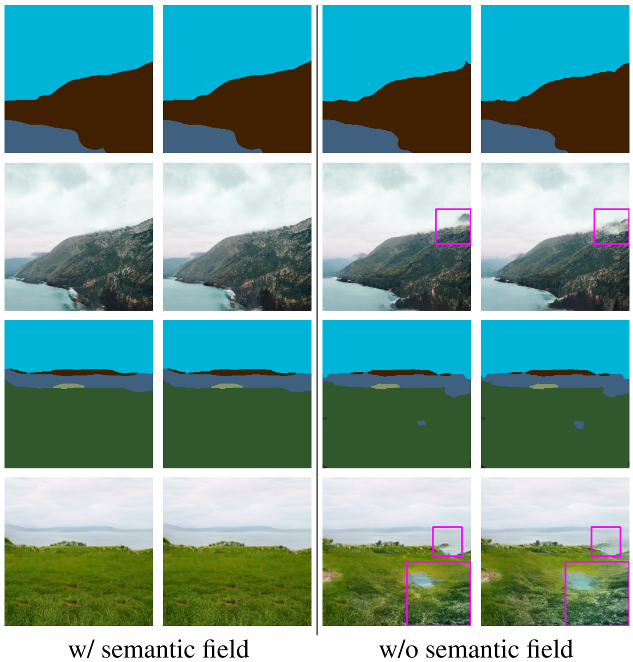

Figure 4. **The effectiveness of semantic field.** The cropped patch clearly indicates the minor change in the semantic masks across different viewpoints (the first and second columns are adjacent viewpoints) brings the unwanted large region change in RGB images generated by SPADE.

In addition, we use depth maps to learn the geometry of the semantic field. In detail, the infilled semantic masks are first converted to RGB images via SPADE, then processed by a monocular depth estimation network [50] to predict depth maps. A scale- and shift-invariant loss [50, 71] is utilized to calculate the difference between the rendered depth map $\mathbf{D}$ and the predicted depth map $\hat{\mathbf{D}}$, which is defined as:

$$\mathcal{L}_{\text{depth}} = \sum_{\mathbf{r}\in\mathcal{R}'} \|(w\mathbf{D}(\mathbf{r}) + q) - \hat{\mathbf{D}}(\mathbf{r})\|^2, \tag{4}$$

where $\mathcal{R}'$ means camera rays of image pixels excluding the sky region. w and q are used to align the scale and shift of $\mathbf{D}$ and $\hat{\mathbf{D}}$, which can be obtained using a least-squares criterion [12, 50].

To separate the foreground and the sky region, we apply a loss on the accumulated transmittance:

$$\mathcal{L}_{\text{trans}} = \sum_{\mathbf{r}\in\mathcal{R}} (\log(T_{\text{fg}}(\mathbf{r})) + \log(1 - T_{\text{fg}}(\mathbf{r}))). \tag{5}$$

This loss enforces the transmittance to be either 0 or 1. The overall loss is described in the supplementary material.

3.2. Natural scene representations

Directly translating multi-view semantic masks obtained in Section 3.1 to RGB images through SPADE fails to produce multi-view consistent rendering, as shown in the supplementary material. To resolve this issue, we learn a natural scene representation to fuse appearance information provided by SPADE. This section describes the generation of natural scenes from the learned semantic field. We first introduce the neural representation of natural scenes. Then, the rendering and training of the scene representation are described.

The geometry of the scene is directly modeled as the trained MLP network f_θ (Eq. 1) of the semantic field. To represent the scene's appearance, we recover an appearance field f_ξ. Following EG3D [6], a tri-plane feature map is adopted to map a point to a feature vector. Specifically, given a point $\mathbf{x}$, it is orthogonally projected to the feature planes to retrieve three feature vectors, which are concatenated into the final feature vector. We use an MLP network to regress the RGB value $\mathbf{c}$ from the aggregated feature vector. The appearance field is defined as:

$$f_\xi : \mathbf{x} \in \mathbb{R}^3 \mapsto \mathbf{c} \in \mathbb{R}^3 \tag{6}$$

For the scene representation, we separately model the foreground and the sky region. The sky is implemented as a 2D image plane generated by a 2D generator network, and we place it at a distance. For a ray classified as 'sky', the sky image plane maps the intersection point (u, v) between the ray and the sky plane to the RGB value.

To efficiently render the scene representation, we leverage the pre-learned scene geometry to guide the sampling of points along camera rays. The mesh is first extracted from the trained MLP network f_θ. Then, for each camera ray, we only predict the color for the point on the mesh surface, similar to [33, 38]. With surface-guided rendering, the computational cost of synthesizing full-resolution images is significantly reduced.

During training, the geometry network f_θ is fixed. The appearance network is optimized based on perceptual and adversarial losses. We first use the learned semantic field to render multi-view semantic masks which are then translated into images using SPADE. The perceptual loss [27] is adopted to compare the rendered image $\mathbf{C}$ and generated image $\hat{\mathbf{C}}$:

$$\mathcal{L}_{\text{feat}}(\hat{\mathbf{C}}, \mathbf{C}) = \left\|\phi(\hat{\mathbf{C}}) - \phi(\mathbf{C})\right\|_2^2, \tag{7}$$

where ϕ denotes the VGG network [57]. Perceptual loss makes training procedure more stable and faster.

Additionally, the adversarial loss is applied to make the rendered images more photorealistic and prevent blurriness caused by the inconsistency of input views, which is demonstrated in GANcraft [18]. Our rendered images are taken as "fake" samples, and generated images are taken as "real" samples. We adopt the OASIS discriminator [61] as our discriminator, and use the same generator and discriminator loss as [61].

Methods	FID↓	KID↓
SPADE [46]+InfiniteNatureZero [32]	149.80	0.080
SPADE [46]+3DPhoto [56]	127.74	0.064
SPADE [46]+AdaMPI [17]	185.96	0.115
GVS* [16]	141.64	0.084
Ours	**109.85**	**0.050**

Table 1. **Quantitative comparisons on the LHQ dataset.** "SPADE + *" means a two-stage pipeline that first generates an image with SPADE and then performs single-view view synthesis. "GVS*" means that we train GVS on the LHD dataset using the strategy in AdaMPI.

4. Experiments

4.1. Datasets

We use the LHQ dataset [58] to train our SPADE and semantic refinement network. The LHQ dataset is a large collection of landscape photos collected from the Internet. To prepare the semantic mask for each image, we use COCO-Stuff [5] to train a DeepLab v2 [8] network. The training data for semantic inpainting is synthesized by the back-warping strategy as mentioned in Section 3.1.

4.2. Implementation details

To train the semantic field, 1024 rays are sampled per learning iteration. For the neural appearance field, we train the neural appearance field at a 256×256 resolution. We train our semantic field and appearance field on 1 NVIDIA RTX 3090 with 24GB of memory. For the semantic field, each model is trained for 48k iterations with batch size 1, which takes approximately 13 hours. For the appearance field, each model is trained for 12k iterations with batch size 4, which takes approximately 3 hours. We use a combination of GAN loss, L2 loss, and perceptual loss for the appearance field. Their weights are 1.0, 10.0, and 10.0, respectively. More training details are described in the supplementary material.

4.3. Metrics

Following Gancraft [18], we use both quantitative and qualitative metrics to evaluate our synthesis methods.

Quantitative metrics. We adopt the widely used metrics, FID [20, 47] and KID [3, 47], to measure the distance between the real and generated distributions. Our experiments are conducted on 6 test scenes. A collection of landscape images from Flickr is obtained to evaluate the quality of our generated images. We make sure that test images are not presented in the LHQ dataset for training. The input semantic mask is obtained by a pretrained semantic segmentation model [8]. For each test scene, we render 330 images using a randomly sampled style code from uniformly sampled camera poses. For both FID and KID metrics, lower values indicate better image quality.

Qualitative metrics. To quantify the aspects that are not addressed by automatic evaluation metrics, we conduct a user study to compare our method with the baseline methods. Specifically, we consider two aspects for human observers: 1) view consistency and 2) photo-realism. 18 visual designers in a 3D content designing company are asked to assign a score on a continuous scale of 1-5 for each aspect per video, where 1 is worse, and 5 is best. Two video sequences of the same scene rendered by two different methods are presented at the same time in random orders. A total of 9 videos per method is rendered for evaluation for each user. For more details about the user study, please refer to the supplementary material.

4.4. Comparisons with baseline methods

We compare our approach with four strong baselines, all of which are evaluated at a 256×256 resolution:

SPADE [46]+3DPhoto [56]. Using layered depth images, 3DPhoto [56] generates free-viewpoint videos from a single image, and it can be trained on the in-the-wild dataset. Following previous works [32, 34, 52] that directly use official pretrained models for evaluation, we also use the official pretrained models for our evaluation. We combine 3DPhoto and SPADE by synthesizing a color image from the semantic mask using SPADE, then applying 3DPhoto to generate a novel view image.

SPADE [46]+AdaMPI [17]. AdaMPI [17] regresses the MPI representation from an image with a network trained on the in-the-wild dataset. We also use official pretrained models for our evaluation. This method can also be combined with SPADE to apply in this setting. We obtain generated images from the given semantic mask using SPADE and then use AdaMPI to synthesize the multi-plane images.

SPADE [46]+InfiniteNatureZero [32]. InfiniteNature-Zero is a model that generates flythrough videos of natural scenes, beginning with a single image. It is trained on the LHQ dataset. We obtain generated images from the given semantic mask using SPADE and then use InfiniteNature-Zero to perform perceptual view generation according to the camera poses used in our test set.

GVS* [16]. We carefully train GVS [16] on the LHQ dataset, using the same training strategy as AdaMPI [56], which takes approximately 2 days. To ensure its style is consistent with other methods, it is finetuned on the test scenes. We find the official inpainting network pretrained by AdaMPI [56] can not deal with large disocclusion regions. To prevent these regions from affecting GVS, GVS is not supervised in these areas. GVS is trained on 3 NVIDIA TITAN A100 40GB graphics cards with 16 batch sizes.

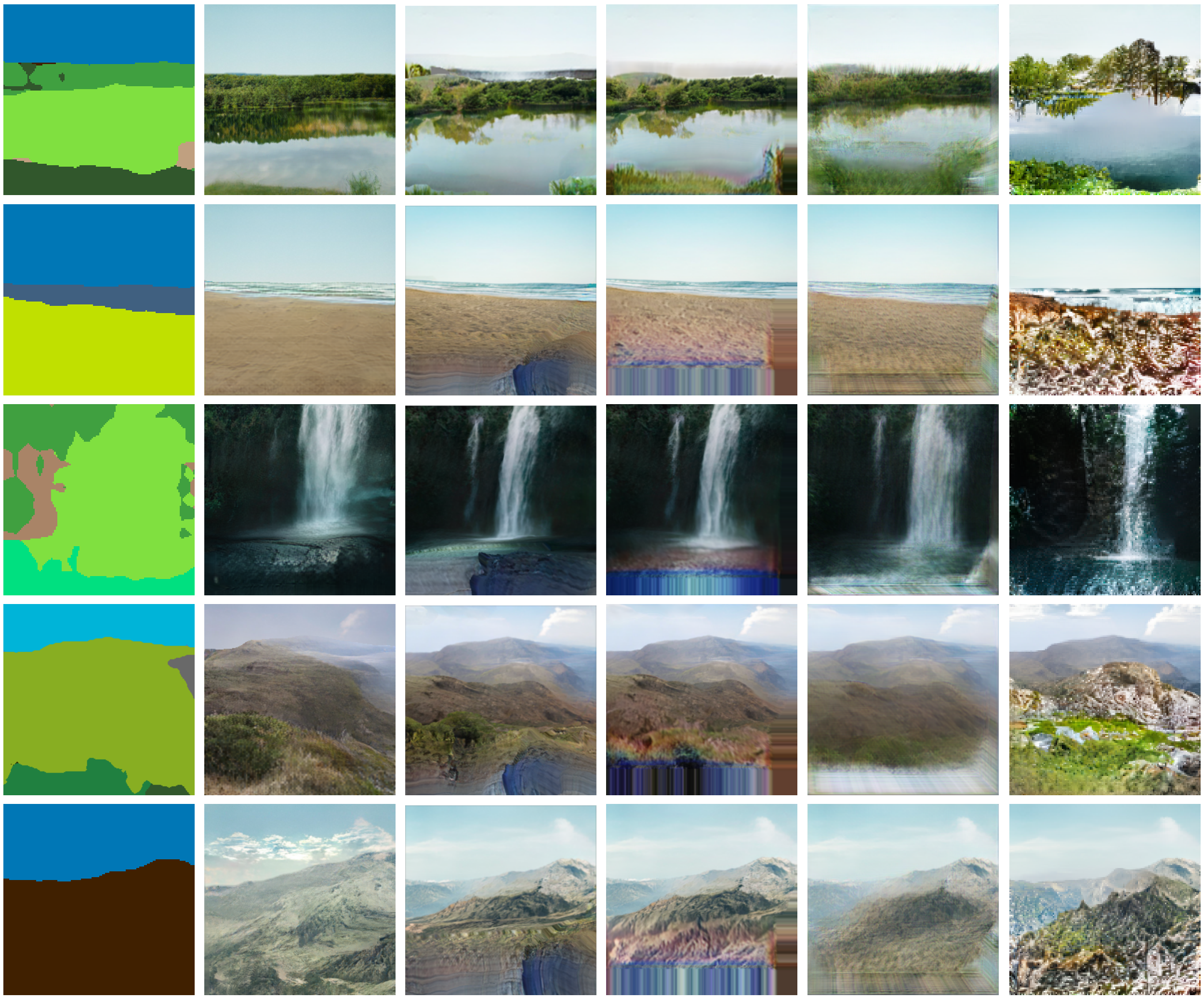

Figure 5. **Qualitative comparisons on the LHQ dataset.** We produce more realistic results compared to all baselines, as demonstrated in the supplementary video.

Fig. 5 shows the results of generated videos of different methods. Directly combining SPADE with 3DPhoto or AdaMPI does not fully utilize the semantic layout information. Therefore, they are prone to erroneously inpainting disocclusion regions. In addition, our approach can render free-viewpoint videos with a wide range of viewpoints, because it is rather easy for our semantic inpainting network to synthesize a large disocclusion region and the appearance information at these regions can be generated via SPADE effortlessly. As GVS* only produces multiple planes at fixed depths, it struggles to represent complex scene layouts for natural images. InfiniteNatureZero [32] is not designed for novel view synthesis and struggles to produce realistic results for the input camera trajectory is different from the training. Besides, InfiniteNatureZero and GVS* use 2D CNNs to generate or refine RGB images, which do not guarantee the inter-view consistency. On the contrary, our method constructs a continuous neural field to fuse appearance information from SPADE, so that more realistic and consistent results can be obtained.

As shown in Table 1, our approach outperforms existing baselines, achieving the smallest FID and KID. Furthermore, Table 2 indicates that users prefer our method and rate our videos the most view-consistent and realistic compared to others.

4.5. Ablation studies

We conduct ablation studies to demonstrate the importance of each component of our method on one test scene. Our key idea is decomposing the task of view synthesis

Methods	Consistency↑	Realism↑
SPADE [46]+InfiniteNatureZero [32]	1.31	2.06
SPADE [46]+AdaMPI [17]	3.63	1.75
SPADE [46]+3DPhoto [56]	3.94	2.19
GVS* [16]	2.63	2.13
Ours	**4.11**	**3.13**

Table 2. **Human preference scores.** Our method achieves the highest photo-realism and multi-view consistency scores according to human raters.

Figure 6. **Qualitative results of ablation studies.** "RGB inpainting" denotes the neural scene representation learned on multi-view images generated by an image inpainting network. This model fails to produce plausible image contents in disocclusion regions. "Ours w/o SF" denotes the scene representation learned on images generated from the (post-inpainting) semantic masks and has the same geometry as our full method. The inconsistency of semantic masks causes large changes in RGB images, resulting in degraded rendering quality.

from a semantic mask into two simpler steps, which first generate multi-view semantic masks and then produce RGB images with SPADE for learning a neural scene representation. To demonstrate the effectiveness of this idea, we design a naive baseline where we use a pre-trained RGB inpainting network to generate multi-view images for recovering the 3D scene. All methods use the same depth map for a fair comparison. Specifically, We generate RGB images by SPADE from the given semantic mask and then use a monocular depth estimation model to predict the depth maps of the generated image. The generated image is then warped to novel views. To infill disocclusion regions, we apply a pretrained RGB inpainting network to images at novel views, which output the final multi-view RGBD images. The pretrained RGB and depth inpainting networks are the official models from AdaMPI [17]. We abbreviate this baseline as "RGB inpainting". As shown in Fig. 6, this model fails to produce photorealistic results under big viewpoint changes. This is because it is challenging for an RGB inpainting network trained on single-view datasets to infill such large missing areas. In contrast, our approach renders high-quality images, which indicates that semantic masks are easier to inpaint than RGB images.

The second important design is our semantic field fusion module, which leverages the semantic field to denoise and fuse semantic masks generated by the inpainting network. To illustrate the necessity of this module, we design the baseline where the infilled semantic masks are directly fed to SPADE to produce RGB images for learning the scene representation. For this baseline, we use the same geometry as our full method. We abbreviate this baseline as "ours w/o SF". The result in Fig. 6 indicates that although this model can synthesize reasonable contents in disocclusion regions, the rendered images tend to be blurry. The reason is that the infilled semantic masks are not view-consistent, especially near semantic edges, resulting in that images generated by SPADE differing significantly across different viewpoints, which makes us difficult to reconstruct the 3D scene. Besides, to demonstrate that our generated multi-view semantic masks are better than those generated by GVS [16], we compare our method with it on the scene used in ablation studies. Following [49], we utilize the average Negative Log-Likelihood (NLL) score to measure the quality of generated semantic masks and the View Semantic Consistency (VSC) score to evaluate the consistency of generated semantic masks. Tab. 3 shows that the quality and consistency of our generated semantic masks outperform those of GVS.

Methods	NLL↓	VSC↓
Ours	**1.60**	**0.049**
GVS	3.33	0.089

Table 3. **Comparison with GVS.** Our multi-view semantic masks are more consistent and have better quality.

5. Conclusion

In this work, we introduce an AI-enabled content creation tool that uses a single semantic mask to produce a 3D scene, which can be rendered from arbitrary viewpoints. Our method only requires single-view image collections for training, without the need for multi-view images. To achieve this, we propose a novel pipeline that first learns an inpainting network to generate novel views of the input semantic mask and then optimizes a 3D semantic field to render view-consistent semantic masks, which are subsequently fed into a 2D generator SPADE to produce RGB images for learning a neural scene representation. Experiments demonstrate that our method can produce satisfactory photorealistic and view-consistent results that significantly outperform baseline methods.

Acknowledgements. This work was supported by NSFC (No. 62172364), the Alibaba Innovation Research program, and the Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 3

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[4] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 2

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 2, 6

[6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 5

[7] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 3

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3, 6

[9] Yuedong Chen, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Sem2nerf: Converting single-view semantic masks to neural radiance fields. *arXiv preprint arXiv:2203.10821*, 2022. 3

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 2

[11] Thibaud Ehret, Roger Marí, and Gabriele Facciolo. Nerf, meet differential geometry! *arXiv preprint arXiv:2206.14938*, 2022. 3

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 5

[13] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv preprint arXiv:2203.15224*, 2022. 3

[14] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4

[15] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 3

[16] Tewodros Amberbir Habtegebrial, V. Jampani, Orazio Gallo, and Didier Stricker. Generative view synthesis: From single-view semantics to novel-view images. *Advances in neural information processing systems*, 2020. 2, 6, 7, 8

[17] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH*, 2022. 2, 3, 6, 7, 8

[18] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14072–14082, 2021. 3, 4, 5, 6

[19] Peter Hedman and Johannes Kopf. Instant 3d photography. *ACM Transactions on Graphics (TOG)*, 37:1 – 12, 2018. 2

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[21] Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12528–12537, 2021. 2

[22] Hsin-Ping Huang, Hung-Yu Tseng, Hsin-Ying Lee, and Jia-Bin Huang. Semantic view synthesis. In *European Conference on Computer Vision*, 2020. 2

[23] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18342–18352, 2022. 3

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 3

[25] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 867–876, 2022. 3

[26] Jaebong Jeong, Janghun Jo, Sunghyun Cho, and Jaesik Park. 3d scene painting via semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2262–2272, 2022. 3

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3, 5

[28] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 3

[29] Johannes Kopf, Kevin Matzen, Suhib Alsisan, Ocean Quigley, Francis Ge, Yangming Chong, Josh Patterson, Jan-Michael Frahm, Shu Wu, Matthew Yu, et al. One shot 3d photography. *ACM Transactions on Graphics (TOG)*, 39(4):76–1, 2020. 2

[30] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. 3

[31] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. 2

[32] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. *ArXiv*, abs/2207.11148, 2022. 2, 6, 7, 8

[33] Haotong Lin, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields with learned depth-guided sampling. *arXiv preprint arXiv:2112.01517*, 2021. 5

[34] Andrew Liu, Richard Tucker, V. Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14438–14447, 2021. 2, 6

[35] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. *arXiv preprint arXiv:2201.07786*, 2022. 3

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[37] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[38] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H Mueller, Chakravarty R Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. Donerf: Towards real-time rendering of compact neural radiance fields using depth oracle networks. In *Computer Graphics Forum*, volume 40, pages 45–59. Wiley Online Library, 2021. 5

[39] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: stylized neural implicit representations for 3d scenes. *arXiv preprint arXiv:2207.02363*, 2022. 3

[40] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 3

[41] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. In *2021 International Conference on 3D Vision (3DV)*, pages 951–961. IEEE, 2021. 3

[42] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 3

[43] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics (TOG)*, 38:1 – 15, 2019. 2

[44] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 3

[45] Byeongjun Park, Hyojun Go, and Changick Kim. Bridging implicit and explicit geometric transformations for single-image view synthesis. *ArXiv*, abs/2209.07105, 2022. 2

[46] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2332–2341, 2019. 1, 3, 6, 8

[47] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[48] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[49] Xiaotian Qiao, Gerhard P Hancke, and Rynson WH Lau. Learning object context for novel-view scene layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16990–16999, 2022. 2, 8

[50] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2, 3, 5

[51] C. Rockwell, David F. Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14084–14093, 2021. 2

[52] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14336–14346, 2021. 2, 6

[53] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware im-

age synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 3

[54] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *arXiv preprint arXiv:2206.07695*, 2022. 3

[55] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 231–242, 1998. 2

[56] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8025–8035, 2020. 2, 3, 6, 7, 8

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[58] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14144–14153, 2021. 2, 3, 6

[59] Jingxiang Sun, Xuan Wang, Yichun Shi, Lizhen Wang, Jue Wang, and Yebin Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *arXiv preprint arXiv:2205.15517*, 2022. 3

[60] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022. 3

[61] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. *arXiv preprint arXiv:2012.04781*, 2020. 1, 3, 5

[62] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. *arXiv preprint arXiv:2209.03494*, 2022. 3

[63] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *European Conference on Computer Vision*, 2018. 2

[64] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 3

[65] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 4

[66] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1, 3

[67] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2020. 2

[68] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. *arXiv preprint arXiv:2204.00928*, 2022. 3

[69] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 4

[70] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2

[71] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 5

[72] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields. In *European Conference on Computer Vision*, pages 717–733. Springer, 2022. 3

[73] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 3

[74] Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5449–5458, 2022. 3

[75] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 3

[76] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ArXiv*, abs/1805.09817, 2018. 2