Semantic 3D-Aware Portrait Synthesis and Manipulation Based on Compositional Neural Radiance Field

Tianxiang Ma\textsuperscript{1,2*}, Bingchuan Li\textsuperscript{3*}, Qian He\textsuperscript{3}, Jing Dong\textsuperscript{2†}, Tieniu Tan\textsuperscript{3,4*},
\textsuperscript{1}School of Artificial Intelligence, University of Chinese Academy of Sciences
\textsuperscript{2}CRIPAC & NLPR, Institute of Automation, Chinese Academy of Sciences
\textsuperscript{3}ByteDance Ltd, Beijing, China
\textsuperscript{4}Nanjing University
\texttt{tianxiang.ma@cripac.ia.ac.cn, \{libingchuan, heqian\}@bytedance.com, \{jdong, tnt\}@nlpr.ia.ac.cn}

Abstract

Recently 3D-aware GAN methods with neural radiance field have developed rapidly. However, current methods model the whole image as an overall neural radiance field, which limits the partial semantic editability of synthetic results. Since NeRF renders an image pixel by pixel, it is possible to split NeRF in the spatial dimension. We propose a Compositional Neural Radiance Field (CNeRF) for semantic 3D-aware portrait synthesis and manipulation. CNeRF divides the image by semantic regions and learns an independent neural radiance field for each region, and finally fuses them and renders the complete image. Thus we can manipulate the synthesized semantic regions independently, while fixing the other parts unchanged. Furthermore, CNeRF is also designed to decouple shape and texture within each semantic region. Compared to state-of-the-art 3D-aware GAN methods, our approach enables fine-grained semantic region manipulation, while maintaining high-quality 3D-consistent synthesis. The ablation studies show the effectiveness of the structure and loss function used by our method. In addition real image inversion and cartoon portrait 3D editing experiments demonstrate the application potential of our method.

Introduction

Photo-realistic image synthesis methods based on 2D Generative Adversarial Networks, such as StyleGAN (Karras, Laine, and Aila 2019; Karras et al. 2020), have achieved widespread success. However, they are designed for single-view image generation and do not have the ability to generate 3D-consistent images.

The 3D-aware GAN methods aim to synthesize 3D-consistent images. Many of these methods (Sitzmann, Zollhöfer, and Wetzstein 2019; Yariv et al. 2020; Mildenhall et al. 2020; Schwarz et al. 2020; Niemeyer and Geiger 2021; Chan et al. 2021) require supervision based on multi-view data or synthesize low-resolution images, which limits the applicability of such methods. Other approaches (Gu et al. 2021; Chan et al. 2022; Or-El et al. 2022) explore 3D-consistent image generation under single-view training data, which are inspired by StyleGAN’s \textit{w} latent code modulation network and design similar modulated neural radiance field network. These methods effectively solves the problem that training relies on multi-view data. But they model the whole image as an overall neural radiance field, which limits the semantic part manipulation of synthetic images. (Sun et al. 2022) introduces semantic segmentation maps into the rendering process, but it still uses a single rendering network and do not explicitly decouple the local semantic parts. It cannot directly edit the semantic regions of the synthetic image by manipulating the \textit{w} latent code.

In this paper, we propose the first compositional neural radiance field for semantic 3D-aware portrait synthesis and manipulation, namely CNeRF. It divides the portrait image by semantic regions and learns an independent neural radiance field for each region. Therefore, we can independently control the local semantic regions of the synthetic 3D portrait by manipulating the latent code of the neural radiance field in each semantic region. Specifically, we split the overall generator network into \textit{k} local semantic 3D generators. Each sub-network consists of modulated MLPs to output feature values, color values, mask values, and residual SDF representation of 3D space for each semantic region. Afterwards, the CNeRF conducts semantic part fusion and volume aggregation to render the complete 2D features, 2D colors and corresponding 2D masks. The 2D features and 2D masks can be further processed by a style-based 2D generator to obtain higher resolution portrait. The whole training process is divided into a low-resolution CNeRF rendering stage and a high-resolution image synthesis stage.

Our method is evaluated mainly on FFHQ dataset and corresponding semantic masks. While generating quality and diversity comparable to state-of-the-arts, our method can also finely manipulate semantic parts of generated portraits, and decouple the shape-texture of the semantic regions. In addition we evaluate our approach on a proprietary cartoon portrait dataset. High-quality generation and semantic part editing of the portraits demonstrate the application potential of our method.

To summarize, our main contributions are as following:

\begin{itemize}
  \item We present the first compositional neural radiance field, which divides the image by semantic regions and learns an independent neural radiance field for each region.
  \item Based on CNeRF, we propose a semantic 3D-aware por-
\end{itemize}
trait synthesis and manipulation method and achieve disentanglement of rendered shapes and textures within each semantic region.

- We propose the global and semantic discriminative loss and shape texture decoupling loss to address semantic entanglement during compositional rendering.
- Experiments reveal that our approach can finely manipulate the semantic parts, while maintaining high-fidelity 3D-consistent portrait synthesis.

Related Work

2D Generative Adversarial Networks

2D GANs (Goodfellow et al. 2014) have achieved remarkable success in computer vision. There are many approaches focused on exploring image generation for specific scenarios, including scene generation (Park et al. 2020; Arad Hudson and Zitnick 2021), face generation (Wiles, Koepke, and Zisserman 2018; Lee et al. 2020), human generation (Ma et al. 2021; Ren et al. 2022), etc. One of the most widely used generative frameworks is StyleGAN (Karras, Laine, and Aila 2019; Karras et al. 2020, 2021), whose latent space provides a way to edit different attributes of the generated images. Some approaches explore the editing methods in StyleGAN’s latent space, such as InterFaceGAN (Shen et al. 2020), StyleFlow (Abdal et al. 2021), etc. Since StyleGAN is a noise-to-image type of generation method, it cannot input real images for editing. GAN inversion methods are proposed to solve the problem of real image editing using pre-trained StyleGAN, for instance, e4e (Tov et al. 2021), and PTI (Roich et al. 2021).

3D-Aware GANs

Recently, Neural Implicit Representations (NIR) are widely used for scene 3D modeling (Chabra et al. 2020; Chibane, Pons-Moll et al. 2020; Jiang et al. 2020), object shape and appearance rendering (Michalkiewicz et al. 2019; Niemeyer et al. 2019; Atzmon and Lipman 2020; Chibane, Alldieck, and Pons-Moll 2020; Gropp et al. 2020). Some methods (Sitzmann, Zollhoefer, and Wetzstein 2019; Mildenhall et al. 2020; Niemeyer et al. 2020) learn neural implicit representation by using multi-view 2D images without 3D data supervision. One of the most important methods is the Neural Radiance Fields (NeRF) (Mildenhall et al. 2020), which represents the 3D scene as a series of neural radiance and density fields and uses volume rendering (Kajiya and Von Herzen 1984) technique for 3D reconstruction. Many approaches recently focused on 3D-aware GAN under single-view image supervision. Pi-GAN (Chan et al. 2021) proposes a siren-based neural Radiance field and uses global latent code to control the generation of shapes and textures. GIRAFFE (Niemeyer and Geiger 2021) uses a two-stage rendering process, first generating a low-resolution feature map with a volume renderer, and then learning to boost the resolution of the output with a 2D CNN network. StyleNeRF (Gu et al. 2021) integrates NeRF into a style-based generator to improve rendering efficiency and 3D-consistency of high-resolution image generation. EG3D (Chan et al. 2022) proposes a tri-plane 3D representation method to improve rendering computational efficiency and generation quality. StyleSDF (Or-El et al. 2022) merges a Signed Distance Fields 3D representation with a style-based 2D generator. Recently, some 3D-aware GAN methods (Sun et al. 2022;
Chen et al. (2022) introduce semantic segmentation into the generative network. But they render images and segmentation masks in single network with insufficient semantic part decoupling, which cannot directly edit the semantic regions. In this paper, our method explicitly decomposes the rendering network into multiple semantic region renderers for better semantic decoupling and semantic region editing.

**Compositional Image Synthesis**

Some methods (Arandjelović and Zisserman 2019; Azadi et al. 2020; Sbai, Couprie, and Aubry 2021) generate complete scenes by combining different images as elements directly at the network input. Other approaches (Eslami et al. 2016; Yang et al. 2017; Burgess et al. 2019; Greff et al. 2019; Ehrhardt et al. 2020; Yang, Chen, and Soatto 2020; Arad Hudson and Zitnick 2021; Shi et al. 2022) explore the independent implicit representations of different objects in the scene image, which allows to manipulate the objects in the generated scene by controlling the latent code of each object. For 3D-aware GAN methods, it is a novel idea to split the rendered image according to semantic regions and then design a compositional volume rendering method, which allows for individual manipulation of the semantic region in the generated 3D-consistent images.

**Methodology**

**Overview**

The overall model is divided into two stages. Firstly, we train a novel compositional neural radiance field to divide the image by semantic regions and learn an independent neural radiance field for each region, and finally fuses them and renders the complete 2D portrait. Secondly, we utilize a style-based 2D generator to upsample the low-resolution 2D images and obtain high-quality 3D-consistent portraits. In the inference, we can achieve semantic part control of the 3D-consistent portraits by manipulating the latent code of each semantic generator in CNeRF. The overall framework is shown in Figure 2, and the details are specified below.
Compositional Neural Radiance Field

We propose the Compositional Neural Radiance Field (CN-erF), as shown on the lower left of Figure 2. We split the overall scene renderer into $k$ local semantic 3D generators, denoted as $G_i (i \in \{1, \ldots, k\})$, with $k$ representing the number of semantic categories we set according to the dataset. For the FFHQ dataset we set 9 local semantic categories, i.e., background, face, eyes, eyebrows, mouth, nose, ears, hair and the area below the head. Each local generator consists of modulated Multilayer Perceptron (mMLP) (Chan et al. 2021) and fully connected layers (FC Layers) of the same structure. The generator $G_i$ takes as input the 3D coordinates $x$, the view direction $v$, and uses the latent code $w_i$ learned from the noise of the standard Gaussian distribution as the network modulation parameters. $G_i$ outputs two view-dependent color value $c_i(x, v)$ and mask value $m_i(x, v)$, where $m_i(x, v)$ is obtained from the feature output $f_i(x, v)$. In addition $G_i$ outputs a view-independent residual SDF value $\Delta d_i(x)$, where SDF stands for Signed Distance Fields, a proxy of the volume density (Yariv et al. 2021; Or-El et al. 2022). The network structure of $G_i$ is shown in the lower right of Figure 2.

Local Semantic 3D Generator. For each local semantic 3D generator $G_i$, we further decouple the shape and texture properties by splitting the mMLP network into two parts to learn the shape and texture of each semantic part separately. To balance the model performance and the number of parameters, our shape network uses a 3-layer mMLP with SIREN activation (Sitzmann et al. 2020) and the texture network uses a 2-layer mMLP. The expression for $j$th layer of the mMLP is as follows,

$$\phi_j(x) = \sin \left( \gamma_j \left( W_j \cdot x + b_j \right) + \beta_j \right),$$

where $W_j$ and $b_j$ are network parameters, $\gamma_j$ and $\beta_j$ are modulation coefficients computed from the input $w_i (w_{\text{shape}}$ or $w_{\text{texture}}$) latent code, and the $3$-layer MLP with LeakyReLU activation. The $w_i$ of each semantic generator may be the same or different at training iteration depending on random settings. Based on the above network, we use the FC layers to output different features at specific layers. Where in addition to the residual SDF value, other outputs are affected by the view direction.

Residual SDF Representation. We tried two types of volume density representation: direct volume density representation and SDF proxy. However, in our experiments we found that splitting the overall generator into local semantic generators destroys the learning process of object 3D shape and leads to incorrect 3D geometric reconstruction. We therefore propose a new volume density proxy, residual SDF representation, for compositional rendering. We add the $\Delta d_i(x)$ output from each semantic generator to a global SDF $d_0(x)$ initialized as a sphere to serve as the true SDF representation for the whole scene. This allows each $G_i$ to learn only the changes of the SDF function value in corresponding region without learning the real physically meaningful SDF value, which makes good 3D geometric reconstruction possible. We transform the residual SDF values of the 3D semantic parts into the density field of the overall scene following the equation below,

$$\sigma(x) = K_\alpha(d(x)) = \frac{1}{\alpha} \cdot \text{Sigmoid} \left( \frac{-d(x)}{\alpha} \right),$$

$$d(x) = d_0(x) + \sum_i \Delta d_i(x), \quad i \in \{1, \ldots, k\},$$

where $\alpha$ is a learnable parameter to regulate the density tightness of the object surface (Or-El et al. 2022).

Semantic-Based Fusion. In addition to learning a good global density representation, we need to fuse the color values output by $G_i$ to form a complete color value. We utilize the idea of semantic mask-based fusion. $G_i$ outputs the color $c_i$ and feature $f_i$ along with the mask $m_i$ for the corresponding region, so we use $m_i$ to perform a weighted fusion of $c_i$ and $f_i$,

$$c(x, v) = \sum_i c_i(x, v) \cdot m_i(x, v), \quad i \in \{1, \ldots, k\},$$

$$f(x, v) = \sum_i f_i(x, v) \cdot m_i(x, v), \quad i \in \{1, \ldots, k\}.$$  

We do not limit the value range of $m_i$ during the semantic fusion, but learn 2D masks end-to-end.

Volume Aggregation. For each pixel, we query the samples on a ray that originates at the camera position $o$, and points at $r(t) = o + tv$. We utilize the classical volume rendering equations to accumulate the samples on the ray,

$$C(r) = \int_{t_a}^{t_f} T(t) \cdot \sigma(r(t)) \cdot c(r(t), v) dt,$$

$$F(r) = \int_{t_a}^{t_f} T(t) \cdot \sigma(r(t)) \cdot f(r(t), v) dt,$$

$$M_i(r) = \int_{t_a}^{t_f} T(t) \cdot \sigma(r(t)) \cdot m_i(r(t), v) dt,$$

where $T(t) = \exp \left( -\int_{t_s}^{t_f} \sigma(r(s)) ds \right)$. We use the discrete form of the volume aggregation method following NeRF (Mildenhall et al. 2020), and obtain the rendered 2D color, 2D feature and $k$ 2D semantic region masks.

Global and Semantic Discriminators

To help the training of our CNeRF model, we propose two discriminators, the Global-Discriminator (GD), which judges the overall rendering result, and the Semantic-Discriminator (SD), which judges the semantic parts. The GD takes as input the rendered 2D mask and 2D color image and outputs the predicted true/false label as well as the estimated view direction of the input generated image. The estimation of view direction is to ensure a good 3D-consistency of the synthesized image. The SD is designed to enhance semantic region disentanglement. In each iteration of training, the model randomly selects one of the $k$ semantic categories and extracts the corresponding region in the 2D color image based on the generated 2D mask. The SD learns to discriminate the reality of the semantic part input and to predict the semantic category of the input image. Two discriminators consist of a residual CNN, with the network divided into two
output layers at the end. The input of GD has two branches, image CNN and mask CNN. Then the output features of the two branches are added together. Besides, the input of SD is a single branch of semantic region image.

**Loss Functions**

**Global and Semantic Discriminative Loss.** For the adversarial losses, we utilize the non-saturating GAN loss with R1 regularization (Mescheder, Geiger, and Nowozin 2018) for our two discriminators, denoted as $L_{adv}^{global}$ and $L_{adv}^{semantic}$. In addition the two discriminators have additional view direction prediction and semantic category prediction, respectively. We use a smoothed $L1$ loss and a cross-entropy loss to constrain these two learning processes, denoted as $L_{view}^{global}$ and $L_{classify}^{semantic}$.

**Shape Texture Decoupling Loss.** Based on our local semantic 3D generator network, we propose a shape texture decoupling loss. It can help better decouple the shape-texture of each semantic region.

$$L_{STD} = \|\mathcal{M}(w_s^1, w_t^1) - \mathcal{M}(w_s^0, w_t^0)\|_1 + \max \{\|\phi(I(w_s^1, w_t^1)) - \phi(I(w_s^0, w_t^0))\|_1$$

$$- \|\phi(I(w_s^1, w_t^2)) - \phi(I(w_s^0, w_t^2))\|_1 + m, 0\},$$

where $w_s$ and $w_t$ are the abbreviation of $w_{shape}$ and $w_{texture}$. The first half of the formula indicates the expectation that the generated 2D masks are consistent when we fix the $w_{shape}$ and change the $w_{texture}$. The $\mathcal{M}(*, *)$ represents the rendered 2D mask. The second half represents a contrast loss, where we expect that inputting the same $w_{texture}$ will make the generated 2D color images closer in feature space than inputting two different $w_{texture}$. The $\phi$ is the middle feature layer of the pre-trained VGG network, and $m$ is a fixed margin.

**Eikonal Loss and Minimal Surface Loss.** Eikonal Loss guarantees that the learned overall SDF of the object is physically valid (Gropp et al. 2020).

$$L_{Eik} = \mathbb{E}_x(\|\nabla d(x)\|_2 - 1)^2.$$  

Minimal Surface Loss is used to penalize the SDF value to ensure that the surface of the 3D object is smooth and free of false and invisible surfaces (Or-El et al. 2022).

$$L_{MS} = \mathbb{E}_x(\exp(-100d(x))).$$

The overall loss function is,

$$L_{overall} = L_{adv}^{global} + L_{view}^{global} + L_{adv}^{semantic} + L_{classify}^{semantic} + L_{STD} + L_{Eik} + L_{MS}.$$  

The weights of each loss item are $\lambda_{view} = 15, \lambda_{Eik} = 0.1, \lambda_{MS} = 0.001$, and the rest are 1.

**High-Resolution Synthesis**

Direct synthesis of high-resolution 3D-consistent images is difficult, so we use a two-stage training approach (Gu et al. 2021; Or-El et al. 2022). First the $64 \times 64$ resolution 2D color, 2D feature map, and 2D masks are rendered. Then the CNeRF network weights are fixed and a style-based 2D generator is used to learn higher resolution image synthesis. We utilize the structure of StyleGAN2 (Karras et al. 2020) with input 2D feature, and add a mask input branch. We use the CNeRF’s latent code $w$ as input to the mapping network of the 2D generator. The 2D generator is shown in the upper right of Figure 2. We equip the 2D generator with a high-resolution discriminator, which has the same function as the Global-Discriminator of the previous stage.

**Experiments**

**Datasets.** We train and evaluate our model mainly on the FFHQ (Karras, Laine, and Aila 2019) dataset, which contains 70,000 high-quality portrait images with a maximum resolution of $1024 \times 1024$. We use the DeeplabV3 (Chen et al. 2017) for semantic segmentation of these portrait images. In addition we utilize a 3D rendering pipeline to produce a high-quality cartoon portraits dataset, which contains 20,000 images at 1024 resolution. Similarly we perform facial semantic segmentation on this dataset. Our experimental re-
Table 1: Quantitative evaluation. All methods are evaluated under FFHQ dataset. The resolution of the synthesized image is shown in parentheses. A: 3D-consistent image synthesis, B: Semantic region manipulation, C: Semantic-level shape texture decoupling.

<table>
<thead>
<tr>
<th>Method</th>
<th>FID</th>
<th>KID</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>PiGAN (256)</td>
<td>83.0</td>
<td>85.8</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StyleNeRF (256)</td>
<td>8.0</td>
<td>3.7</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StyleSDF (256)</td>
<td>11.5</td>
<td>2.6</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EG3D (512)</td>
<td>4.7</td>
<td>0.13</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FENeRF (128)</td>
<td>28.2</td>
<td>17.3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ours (256)</td>
<td>13.2</td>
<td>1.2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ours (512)</td>
<td>12.3</td>
<td>0.21</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

First, we train CNeRF using 64×64 resolution portrait images and segmentation masks, and render each output at 64×64 resolution. We use ADAM (Kingma and Ba 2014) optimizer with learning rates of $2 \times 10^{-5}$ and $2 \times 10^{-4}$ for the generator and discriminator respectively and $\beta_1 = 0, \beta_2 = 0.9$. Second, we freeze the CNeRF weights and train the 2D generator with same setup to StyleGAN2. Our experiments are carried out on 4 32GB Tesla V100 GPUs. Due to limited space more implementation details and detailed architecture of 2D generator and discriminators as well as more experimental results are presented in the supplementary material.

Comparisons

3D-Consistent Portrait Synthesis. Figure 3 shows the comparative results of 3D-consistent portrait generation. Our method is able to accurately reconstruct the 3D shape of the faces and also generate 2D masks corresponding to the faces. Our face 3D shape reconstruction is comparable to StyleSDF, but our method has semantic part manipulation ability that other comparison methods do not have. Table 1 presents the quantitative comparison results of our method with SOTA 3D-aware GAN methods, and the performance of our method is comparable to that of StyleSDF. The main contribution of our approach, however, is to synthesize 3D-consistent images while achieving good semantic-level property manipulation.

Semantic Region Manipulation. Figure 1 visually demonstrates the manipulation of the semantic parts of the synthesized faces, while ensuring 3D-consistent image synthesis. Our method can also control the shape and texture of the semantic regions separately, such as hair region in the figure. In addition figure 4 shows the comparison results of our method and FENeRF in real image semantic part manipulation. We use PTI (Roich et al. 2021) to train GAN inversion. Our method inverts more realistic face images.
with better semantic part controllability. For example, our method can control the position as well as the shape of the eyebrows, while FENeRF can only edit at the original eyebrow area. Our model is able to decouple semantic region shape and texture properties, such as the displayed hair area, and change the background, which FENeRF cannot do. In addition our method manipulates the semantic parts by directly controlling the $w_i$ latent code of each semantic generator, while FENeRF can only edit the mask manually and perform semantic editing with the help of GAN inversion.

Applications
Our method can be applied to the editing of real images, as shown in Figure 4. The quality of the image inversion and the semantic part manipulation effect illustrate the great potential of our method for application. In addition we test our method on a proprietary cartoon portrait dataset, as shown in Figure 5. High-quality 3D-consistent portrait synthesis as well as accurate and diverse semantic part editing demonstrate the potential of our approach to 3D animation, digital human and metaverse applications.

Ablation Studies
We conduct ablation experiments on our CNeRF model from two aspects. For the overall 3D-consistent image synthesis ability, the residual SDF representation proposed in this paper plays a key role for compositional neural radiance field, as shown in Figure 6. When using the original volume density or the direct SDF proxy, our CNeRF produces incorrect 3D reconstruction and the synthesized images lose 3D-consistency. For the semantic part decoupling capability, we perform ablation experiments for the two correlated loss functions proposed in this paper, as shown in Figure 7. When using semantic discriminative loss (and Semantic-Discriminator), there is more accurate control over the semantic parts. When using shape texture decoupling loss, the model’s performance is improved on the disentanglement of shape and texture properties and editing diversity.

Conclusion
In this paper, we propose a novel Compositional Neural Radiance Field for semantic 3D-aware portrait synthesis and manipulation. It enables high-quality 3D-consistent image synthesis while achieving independent manipulation of the semantic parts and decoupling the shape and texture within each semantic region. Qualitative and quantitative comparison experiments demonstrate the effectiveness and novelty of our method. The application performance of real image inversion and cartoon portrait 3D-aware editing show the potential of the method. There is still room to improve the quality of our approach in 3D reconstruction, and in future work we will explore the combination of our method with the latest 3D-aware GANs, such as EG3D.
Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant No. 2021YFC3320103.

References


