FENeRF: Face Editing in Neural Radiance Fields

Jingxiang Sun\textsuperscript{1}\* Xuan Wang\textsuperscript{2}\† Yong Zhang\textsuperscript{2} Xiaoyu Li\textsuperscript{2}
Qi Zhang\textsuperscript{2} Yebin Liu\textsuperscript{3} Jue Wang\textsuperscript{2}

\textsuperscript{1}University of Illinois at Urbana-Champaign \textsuperscript{2}Tencent AI Lab \textsuperscript{3}Tsinghua University

Abstract

Previous portrait image generation methods roughly fall into two categories: 2D GANs and 3D-aware GANs. 2D GANs can generate high fidelity portraits but with low view consistency. 3D-aware GAN methods can maintain view consistency but their generated images are not locally editable. To overcome these limitations, we propose FENeRF, a 3D-aware generator that can produce view-consistent and locally-editable portrait images. Our method uses two decoupled latent codes to generate corresponding facial semantics and texture in a spatial-aligned 3D volume with shared geometry. Benefiting from such underlying 3D representation, FENeRF can jointly render the boundary-aligned image and semantic mask and use the semantic mask to edit the 3D volume via GAN inversion. We further show such 3D representation can be learned from widely available monocular image and semantic mask pairs. Moreover, we reveal that joint learning semantics and texture helps to generate finer geometry. Our experiments demonstrate that FENeRF outperforms state-of-the-art methods in various face editing tasks. Code is available at https://github.com/MrTornado24/FENeRF.

1. Introduction

Photo-realistic image synthesis via Generative Adversarial Networks is an important problem in computer vision and graphics. Specifically, synthesizing high-fidelity and editable portrait images has gained considerable attention in recent years. Two main classes of methods have been proposed: 2D GAN image generation and 3D-aware image synthesis techniques.

Despite their great success of synthesizing highly-realistic and even locally-editable images, 2D GAN methods all ignore the projection or rendering process of the underlying 3D scene, which is essential for view consistency. Consequently, they produce inevitable artifacts when changing the viewpoint of generated portraits. In order to overcome this issue, the Neural Radiance Fields (NeRF) \cite{mildenhall2020nerf} have been explored to develop 3D-aware image synthesis techniques. Some of these methods \cite{niemeyer2020cips, yang2019cips} adopt vanilla NeRF generator to synthesize free-view portraits that are not editable, and the results could be blurry. Niemeyer et al. \cite{niemeyer2020cips} employ volumetric rendering techniques to first produce view-consistency 2D feature maps, and then use an additional 2D decoder to obtain the final highly-realistic images. Nevertheless, such method suffers from additional view-dependent artifacts introduced by 2D convolutions and the mirror symmetry problem. To this end, CIPS-3D \cite{yang2019cips} replaces the 2D convolutions with implicit neural representation (INR) network. Unfortunately, all existing 3D-aware GANs do not support the interactive local editing on the generated free-view portraits.

In this paper, we propose a generator that can produce strictly view-consistent portraits, while supports interactive local editing. We adopt the noise-to-volume scheme. The generator takes as input the decoupled shape and texture

\* Work done during an internship at Tencent AI Lab.
† Corresponding Author.
latent code, and generates a 3D volume where the facial semantics and texture are spatially-aligned via the shared geometry. As a learnable 3D positional feature embedding is exploited while generating the texture volume, more details are preserved in the synthesized portraits.

Directly learning this 3D volume representation is challenging due to the absent of suitable, large-scale 3D training data. A possible solution is to use multi-view images [4]. Nonetheless, the inadequate training data harms the representation ability of the 3D semantic volume. To overcome this issue, we make the use of monocular images with paired semantic masks, which are vastly available. Specifically, color and semantic discriminators are employed to supervise the training of the NeRF generator. The color discriminator focuses on image details hence improves the image fidelity. The semantic discriminator takes as input a pair of image and semantic map to enforce the alignment of corresponding content in the 3D volume. Thanks to the spatial-aligned 3D representation, we can use the semantic map to locally and flexibly edit the 3D volume via GAN inversion. In addition, an insight here is that learning the semantic and texture representations simultaneously helps to generate more accurate 3D geometry.

To illustrate the effectiveness of the proposed method, we perform the evaluation on two widely-used public datasets: CelebAMask-HQ and FFHQ. As shown in the experiments, the FENeRF generator outperforms state-of-the-art methods in several aspects. In addition, it supports various downstream tasks. To summarize, our main contributions are as following:

- We present the first portrait image generator that is locally editable and strictly view-consistent, benefiting from the 3D representation in which the semantics, geometry and texture are spatially-aligned.
- We train the generator with paired monocular images and semantic maps without the requirement of multi-view or 3D data. This ensures data diversity and enhances the representation ability of the generator.
- In experiments, we reveal that joint learning the semantic and texture volume can help to generate the finer 3D geometry.

2. Related work

Neural Implicit Representations. Recently Neural implicit scene representation boosts various 3D perception tasks, such as 3D reconstruction and novel view synthesis, with its space continuity and memory efficiency. [33,34,39] represent either scene, objects as occupancy fields or signed distance functions, and 3D data is required for supervision. [35] models scene as neural radiance fields which are baked in weights of MLPs. With differentiable numerical integration of volume rendering, NeRF can be trained on only posed images. Various follow-ups extend NeRF to faster training and testing [8,17,43,49], pose-free [26,31], dynamic scenes [5,50] and animating avatars [15,27,41]. [52] extends NeRF with a semantic segmentation renderer and boosts performance of semantic interpretation. In this work, we build a generative semantic field aligned with neural radiance field. Instead of focusing on scene semantic understanding, we utilize the spatial alignment of facial texture and semantics to achieve semantic-guided attribute editing.

EditNeRF [29] edits generic objects vis propagating 2D scribbles to the 3D space, nevertheless, only supports simple modifications on shape and color, while FENeRF tackles the complex facial deformation and detailed appearance texture in the neural radiance fields.

Face Image Editing with 2D GANs. Generative Adversarial Networks (GANs) are widely used in photo-realistic face editing. Inspired by image-to-image translation, conditional GANs take as condition semantic masks [11,19,40,55] or hand-written sketches [6,7,25] for the interactive editing of face images. SPADE [40] utilizes efficient spatially-adaptive normalization to synthesize photorealistic face images given input semantic layouts. SEAN [55] further enables semantic region-based styling and more flexible facial editing.

In order to provide explicit control of 3D interpretable semantic parameters, SofGAN proposes a semantic occupancy field to render view-consistent semantic maps which provide geometry constraints on image synthesis. However, SofGAN still lacks the interpretation of 3D geometry and considerable semantic labelled 3D scans are required for training semantic rendering. Instead, our FENeRF is trained end-to-end in an adversarial manner without any 3D data or multi-view images. Moreover, we show that our semantic rendering has better view consistency.

3D-Aware Image Synthesis. Despite the tremendous breakthroughs in image generation by deep adversarial models [9,19,21,22,40,46,55], those methods mainly manipulate shape and textures in 2D space without understanding of 3D nature of objects and scenes, resulting in limited pose control ability. To this end, 3D image synthesis methods lift image generation into 3D with explicit camera control. Early approaches [12,36,54] utilize explicit voxel or volume representation thus are limited in resolution. Recently neural implicit scene representations are integrated to generative adversarial models and enable better memory efficiency and multi-view consistency [2,3,10,14,37,38,44,53]. In particular, π-GAN [3] presents siren-based neural radiance field conditioned on a global latent code which entangles geometry and texture. GRAF [44] and Giraffe [37] enable disentangled control of texture and geometry, however, in a global level. The con-
current works [2, 10, 14, 38, 53] demonstrate impressive image generation quality but still don’t support user-interacted local editing. By contrast, our FENeRF enables both global independent styling on texture and geometry as well as local facial attribute editing while preserving view consistency.

3. Method

3.1. Locally Editable NeRF Generator

Our goal is to enable semantic-guided facial editing in 3D space. The main challenges are: 1) We need to decouple shape and texture during image generation. 2) Semantic map has to be strictly aligned with geometry and texture in 3D space. To this end, FENeRF exploits two separate latent codes. The shape latent code is to control the geometry and semantics. The texture code controls the appearance in the texture volume. Moreover, we exploits the three-head architecture in presented generator to individually encode the semantics and texture which are aligned with the underlying geometry depicted in the density volume. We formulate our generator as follows:

\[
G : (\mathbf{x}, \mathbf{d}, \mathbf{z}_s, \mathbf{z}_t, \mathbf{e}_{\text{coord}}) \mapsto (\sigma, \mathbf{c}, \mathbf{s}),
\]

As illustrated in Fig. 2, the proposed generator is parameterized as Multi-Layer Perceptrons (MLPs), which takes as input the 3D point coordinates \( \mathbf{x} = (x, y, z) \), viewing direction \( \mathbf{d} = (\theta, \phi) \) and the learned positional feature embedding \( \mathbf{e}_{\text{coord}} \). Then generates the view-invariant density \( \sigma \in \mathbb{R}^+ \) and semantic labels \( \mathbf{s} \in \mathbb{R}^k \) conditioning on shape latent code \( \mathbf{z}_s \), as well as the view-dependent colour \( \mathbf{c} \in \mathbb{R}^3 \) conditioning on texture code \( \mathbf{z}_t \).

We also utilize a mapping network to map sampled codes into an intermediate latent space \( \mathcal{W} \) and output frequencies \( \gamma \) and phase shifts \( \beta \), controlling the generator through feature-wise linear modulation as done in \cite{3, 42}:

\[
\Phi(x) = W_n(\phi_{n-1} \circ \phi_{n-2} \circ ... \circ \phi_0)(x) + b_n
\]

\[
x_i \mapsto \phi_i(x_i) = \sin(\gamma_i \cdot (W_i x_i + b_i) + \beta_i),
\]

where \( \phi_i : \mathbb{R}^{M_i} \mapsto \mathbb{R}^{N_i} \) is the \( i^{th} \) layer of the network. The input \( \mathbf{x}_i \in \mathbb{R}^{M_i} \) is transformed by the weight matrix \( \mathbf{W}_i \in \mathbb{R}^{N_i \times M_i} \) and the biases \( \mathbf{b}_i \in \mathbb{R}^{M_i} \), and then modulated by the sine nonlinearity.

Nevertheless, utilising the siren-based network only generates images lacking details. Therefore, we introduce a learnable 3D feature grid to compensate high-frequency image details. Specifically, to predict the color for a 3D point \( \mathbf{x} \) with 2D view direction \( \mathbf{d} \), we sample a local feature vector \( \mathbf{e}_{\text{coord}}^\mathbf{x} \) from the feature grid by bi-cubic interpolation, then it is fed into the color branch as additional input. As shown in Fig. 11, it helps to preserve finer-grained image details.

Once the semantic, density and colour fields are generated, we can render those into semantic map and portrait image from arbitrary camera poses via volume rendering. For each 3D point, we first query its color \( \mathbf{c} \), semantic labels \( \mathbf{s} \) and volume density \( \sigma \). To obtain the pixel color \( \mathbf{C}_r \) and semantic label probabilities \( \mathbf{S}_r \), the values of all the samples in the ray are accumulated using the classical volume rendering process. The rendering equations are as follows:
a prior according to [3, 37, 44]. The camera positions are camera pose distribution as gaussian and set pose range as \( p \). Training loss with sampled ones.

This is to encourage the alignment of the face appearance and semantic maps \( D \) discriminates the fidelity of generated portraits. Semantic parameterization of CNN with leaky ReLU activation [20].

In order to learn the unsupervised 3D representations, we design two discriminators \( D_c \) and \( D_s \) both of which are parameterized as CNN with leaky ReLU activation [20]. \( D_c \) discriminates the fidelity of generated portraits. Semantic masks, in addition to face images, are taken as input to \( D_s \). This is to encourage the alignment of the face appearance and semantics. Moreover, we append two channels of \( D_s \) to predict camera pose and then apply camera pose correction loss with sampled ones.

\[
C(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), d)dt, \quad (4)
\]
\[
S(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))s(r(t), d)dt, \quad (5)
\]

where \( T(t) = \exp(-\int_{t_n}^{t} \sigma(r(s))ds) \). In practice, we approximate Eq. 4 and Eq. 5 in a discretized form following NeRF [35]. Note that the three branches of semantics, density and texture share the same intermediate features, and the output density is shared in the processes of colour and semantic rendering as well, to ensure that the generated semantics, density and texture are exactly aligned in 3D space.

3.2. Discriminators

In order to learn the unsupervised 3D representations, we design two discriminators \( D_c \) and \( D_s \) both of which are parameterized as CNN with leaky ReLU activation [20]. \( D_c \) discriminates the fidelity of generated portraits. Semantic masks, in addition to face images, are taken as input to \( D_s \). This is to encourage the alignment of the face appearance and semantics. Moreover, we append two channels of \( D_s \) to predict camera pose and then apply camera pose correction loss with sampled ones.

3.3. Training

During training, we randomly sample camera poses \( \xi \sim p_\xi \) and latent codes \( z_s, z_t \sim \mathcal{N}(0, I) \). We approximate the camera pose distribution as gaussian and set pose range as a prior according to [3, 37, 44]. The camera positions are sampled on the surface of an object-centered sphere, and the camera is always directed towards the origin.

Our training loss is composed of three parts:

\[
\mathcal{L}_{D_c} = \mathbb{E}_{z_s, z_t \sim \mathcal{N}, \xi \sim p_\xi}[\mathcal{L}(D_c(x_s))] + \mathbb{E}_{I \sim \mathcal{P}[I]}[\mathcal{L}(D_c(I))] + \lambda_c ||\nabla D_c(I)||^2 \quad (6)
\]
\[
\mathcal{L}_{D_s} = \mathbb{E}_{z_s, z_t \sim \mathcal{N}, \xi \sim p_\xi}[\mathcal{L}(D_s(x_s, x_c))] + \mathbb{E}_{I \sim \mathcal{P}[I], L \sim \mathcal{L}[L]}[\mathcal{L}(D_s(L, I))] + \lambda_s ||\nabla D_s(L, I)||^2 \quad (7)
\]
\[
\mathcal{L}_{G} = \mathbb{E}_{z_s, z_t \sim \mathcal{N}, \xi \sim p_\xi}[\mathcal{L}(D_c(x_s))] + \mathbb{E}_{z_s, z_t \sim \mathcal{N}, \xi \sim p_\xi}[\mathcal{L}(D_s(x_s, x_c))] + \lambda_p ||\xi - \hat{\xi}|| \quad (8)
\]

where \( f(t) = -\log(1 + \exp(-t)) \), \( \lambda_c, \lambda_s, \lambda_p = 10 \), and \( p_s, p_t \) indicate the distributions of real images \( I \) and semantic maps \( L \) in datasets. The objectives of the image discriminator \( D_c \), semantic discriminator \( D_s \) and generator \( G \) to minimize \( \mathcal{L}_{D_c}, \mathcal{L}_{D_s}, \) and \( \mathcal{L}_{G} \), respectively. \( \mathcal{L}_{D_s} \) shown in Eq. 7 is utilised to discriminate the paired image and semantic map and enforce their spatial alignment. While training generator \( G \) with \( \mathcal{L}_{G} \), we stop gradient back propagation from \( D_s \) into color branch since the gradient would enforce texture to match semantics and lead to loss of fine image details. We adopt non-saturating GAN loss and \( R_1 \) gradient penalty [32]. Moreover, we apply camera pose correction loss (the last term of Eq. 8) to penalize the distance between camera pose \( \hat{\xi}, \xi \) which is fed into generator and predicted by discriminator, respectively. This loss enforces all 3D faces lie in the same canonical pose and encourages a reliable 3D face geometry in avoid of pose drift.

In summary, our method builds a generative implicit representation which encodes facial geometry, texture and semantics jointly in a spatial aligned 3D volume. We further
introduce a learnable feature grid for fine-grained image detail. An auxiliary discriminator further enforces this alignment by taking as input paired synthesized image and semantic map. Furthermore, we notice that semantic rendering significantly improves the quality of synthesized facial geometry as shown in Fig. 3.

4. Experiments

Datasets. We consider two datasets in our experiments for evaluation: CelebAMask-HQ [23] and FFHQ [21]. CelebAMask-HQ contains 30,000 high-resolution face images from CelebA [30] and each image has a segmentation mask of facial attributes. The masks have resolution of $512 \times 512$ and 19 classes including skin, eyebrows, ears, mouth, lip, etc. FFHQ contains 70,000 high-quality face images and we label the semantic classes by BiseNet [51].

Baselines. We compare our model on image synthesis quality with three recent works for 3D-aware image synthesis: GRAF [44], pi-GAN [3] and Giraffe [37]. We also compare the performance of semantic rendering with SofGAN [4] which learns the semantic field with multi-view data and edits the portrait in 2D image plane. For the view consistency of inversion, we compare with InterfaceGAN [45] and E4E [47]. We perform the evaluation of these approaches by leveraging their official implementations.

Implementation details. Our semantic radiance field is parameterized as MLPs with FiLM-conditioned SIREN layers [3]. For discriminator, CoordConv layers [28], residual connections [16] are utilized. During training, we initialize the learning rate to $6 \times 10^{-5}$ for generator $G$, $2 \times 10^{-4}$ for image discriminator $D_c$ and $1 \times 10^{-4}$ for segmentation discriminator $D_s$. We use the Adam optimizer with $\beta_1 = 0, \beta_2 = 0.9$. We start training at $32 \times 32$ with a batch size of 40. Then the resolution is increased to $128 \times 128$ with a batch size of 24. Please refer to the supplemental material for more details.

4.1. Comparisons

Quality evaluation of synthesized image. We conduct quantitative comparisons of synthesized images with the state-of-the-art 3D-aware GAN methods and the results are shown in Table 1. Frechet Inception Distance (FID) [18] and Kernel Inception Distance [1] are used to evaluate the image quality. For fair comparison, we retrain all models on these two datasets with full images (30k for CelebA-HQ and 70k for FFHQ) at $128 \times 128$ resolution. FID score is calculated 2048 randomly sampled images. We reach the state-of-the-art performance on both datasets with FID and KID. This improvement is attributed to: 1) the joint learning with semantic field provides reliable semantic enforcement and facilitate training converge; 2) our learnable feature grid brings high frequency details to synthesized images.

Rendering performance of Semantic field. In our framework, we construct a neural semantic radiance field for a generative 3D face and render both semantic maps and images from arbitrary view points. To evaluate the per-
Figure 7. Global style editing. Style mixing (on left part): Given the Source 1 image and a sequence of source 2 images (on top row), the mixed face image in second row retains texture from source 1 and shape from source 2. Inversely, we show the mixed results with shape of source 1 and texture of source 2 sequence on bottom. Style transfer (on the right part): Given source and target images, we can also transfer the texture style of source image into the target one, and the free-view portraits are generated while facial geometry is preserved.

Table 1. Quantitative comparison of our approach with other 3D-aware GAN methods. Our method outperforms other methods in terms of FID and KID on both CelebA-HQ and FFHQ datasets.

<table>
<thead>
<tr>
<th></th>
<th>FID↓</th>
<th>KID (×10^5)↓</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CelebA-HQ</td>
<td>FFHQ</td>
</tr>
<tr>
<td>GRAF</td>
<td>34.7</td>
<td>66.5</td>
</tr>
<tr>
<td>pi-GAN</td>
<td>14.7</td>
<td>40.3</td>
</tr>
<tr>
<td>Giraffe</td>
<td>16.2</td>
<td>31.9</td>
</tr>
<tr>
<td>FENeRF</td>
<td><strong>12.1</strong></td>
<td><strong>28.2</strong></td>
</tr>
</tbody>
</table>

Performance of semantic rendering, we compare with SofGAN [4] which trains a generative semantic occupancy field supervised by labeled multi-view semantic maps. As Fig. 5 shows, SofGAN is prone to wrong semantic classes at extreme poses and causes artifacts in synthesized face. We attribute the semantic inconsistency of SofGAN to that its semantic rendering relies on the surface construction which is, however, highly ambiguous from pure semantic maps even with multi-view observations. By contrast, FENeRF learns accurate geometry (Fig. 3) benefiting from the joint semantic and image rendering thus keeps view consistency at such challenging poses. Besides, for SofGAN, though the semantic map keeps consistent during zooming in, its synthesised images are not view consistent (e.g. eyes look into different directions). By contrast, FENeRF guarantees strict pixel-level view consistency of synthesised images.

To further explore the inversion ability of semantic rendering, we randomly collect 1000 real portrait images and reproject them into the geometry latent space to recover a semantic field. The left chart in Fig. 4 illustrates that the average mIoU score of these 1000 portraits reaches 0.5 by 100 iterations and finally converges to over 0.7 within 200 iterations. This indicates the geometry latent space of FENeRF covers various portrait shapes. From the visualized example we can see that the facial semantics are reconstructed accurately after 2000 iterations with texture-aligned region boundaries and view consistency.

4.2. Applications

**Disentangled morphing and style mixing.** Our method enables disentangled control on geometry and texture through independent latent sampling. Fig. 6 demonstrates the disentangled morphing along two independent directions. Specifically, we sample two sets of texture and shape codes \((Z_t^1, Z_s^1), (Z_t^2, Z_s^2)\) which synthesize images \(I^1, I^2\) at left bottom and right top corners of Fig. 6. Then we perform linear interpolation in these two latent directions and group them for image synthesis. Our approach also supports style mixing demonstrated in Fig. 7 (on left), proving the effectiveness of our disentangled facial representation.

**3D inversion and style transfer.** Inversion is a popular application of 2D GANs and we further lift it to 3D space and render in arbitrary poses. Though some recent conditional GANs [24,45,47,48] also support pose rotation after inversion, however, suffer from inconsistent identity and texture during rotation. Fig. 9 compares FENeRF with InterfaceGAN and E4E on realistic inversion with pose rotations. InterfaceGAN reconstructs realistic synthesized image for input view but generates obvious artifacts (e.g. sticking texture), while E4E generates inconsistent shape during camera rotations. Compared with 2D GANs, FENeRF outperforms these methods on view consistency of identity and texture with a slight drop of texture details. This is because we reproject the 2D image into a latent space which bounds to a 3D generative volume rather than a 2D space. Therefore, the latent code controls facial properties independent of camera pose. We also support face style transfer which is shown in Fig. 7. Style transfer is more challenging than in-
Interactive image manipulation. Our method enables interactive image editing with semantic guidance. Given the source image, we invert it into texture and shape codes and obtain the inversion results. We manipulate facial attributes on the semantic map (e.g., haircut, eyes, nose and face shape, etc.) and leverage the GAN inversion again. We obtain the corresponding modified free-view portraits.

View consistency of inversion. We compare FENeRF against SOTA methods, InterFaceGAN [45] and E4E [47]. Obviously, when generated portraits rotate, InterFaceGAN suffers from the serious artifacts and E4E fails to preserve the face identity. Moreover, the results of style transfer holds consistency at poses far from the input views.

3D local editing. Both style mixing and transfer manipulate face in a global manner and prove our powerful disentangled control of global texture and geometry. We further find that the semantic field would encourage the latent spaces of geometry and texture to be regional disentangled based on the spatial-aligned 3D volume. To prove that, we take facial attributes manipulation by editing semantic maps interactively. Fig. 8 demonstrates the results of facial attributes editing. We first inverse a real face image into a reconstructed image and the accompanied semantic map. Then we edit this semantic map and reproject it into geometry latent space by optimizing the shape code. Note FENeRF is capable of manipulating facial attributes even with significant deformation (e.g., the enlarged and shorten noses in Fig. 8) while keeping other regions consistent in shape and texture. Moreover, the edited faces by FENeRF still hold strict view consistency.
4.3. Ablation Studies

Benefits of our joint framework. Recall that FENeRF renders face images accompanied by semantic maps. We conduct experiments to explore if this joint framework benefits two kinds of rendering. We first train a FeNeRF without image rendering. Therefore, given a 3D point, we only query its density and semantic labels with rendering a single semantic map. In Fig. 10, FENeRF without image rendering in (a) fails to specify facial regions or converge to reliable geometry. By contrast, training with image rendering in (b) solve these problems. We further explore effect of semantic rendering on geometry quality. (c) and (d) in Fig. 10 demonstrates that semantic rendering encourages the generative surface to smooth and accurate. This result is consistent with Fig. 3.

Effects of learnable coordinate embeddings. To produce fine-grained image details, we introduce a learnable feature grid for the local sampling of coordinate embedding ($e_{\text{coord}}$). We further conduct experiments to explore the effect of $e_{\text{coord}}$ and its injecting positions. As shown in Fig. 11, FENeRF w/o $e_{\text{coord}}$ in (a) generates blurry teeth. Injecting $e_{\text{coord}}$ as in (b) generates sharper details but suffers from artifacts on synthesised images and semantic map since $e_{\text{coord}}$ brings high frequency signal into the volume density when injecting at the beginning of MLP with coordinates. (c) injects $e_{\text{coord}}$ into the color branch and enables both high-frequency image and smooth semantics.

5. Discussions

Limitations. One known limitation is our generator cannot produce HD portrait images due to the computationally-expensive ray casting and volume integration. Besides, GAN inversion is an effective method to perform the local editing of 3D volume. The iterative optimization of inversion, however, is inefficient. As a result, the real-time free-view portrait editing is still an open problem.

Potential Social Impact. FENeRF can be utilized to drive an avatar by changing semantic maps and camera poses to make fake videos. Therefore there are certain risks of fooling face recognition systems, such as vivo detection, with our synthesised fake videos thus it should be careful to deploy the technology.

Conclusion. In this paper we present the first locally editable 3D-aware face generator FENeRF based on implicit scene representation. For using semantic map as editing interface, we introduce a semantic radiance field which aligns facial semantics and texture implicitly in 3D space through the shared geometry. We show that FENeRF realizes fancy applications including style mixing, style transfer, facial attribute editing, and we further push them to a 3D free-view manner with explicit camera control. We hope our work would introduce a promising research direction of editable 3D-aware generative network. For future works, we plan to increase the resolution of synthesised free-view portraits and study the specific 3D-aware GAN inversion as well.

Acknowledgements

This paper is sponsored by NSFC No. 62125107.
References


