Self-supervised Matting-specific Portrait Enhancement and Generation
Yangyang Xu, Zeyang Zhou, and Shengfeng He, Senior Member, IEEE

Abstract—We resolve the ill-posed alpha matting problem from a completely different perspective. Given an input portrait image, instead of estimating the corresponding alpha matte, we focus on the other end, to subtly enhance this input so that the alpha matte can be easily estimated by any existing matting models. This is accomplished by exploring the latent space of GAN models. It is demonstrated that interpretable directions can be found in the latent space and they correspond to semantic image transformations. We further explore this property in alpha matting. Particularly, we invert an input portrait into the latent code of StyleGAN, and our aim is to discover whether there is an enhanced version in the latent space which is more compatible with a reference matting model. We optimize multi-scale latent vectors in the latent spaces under four tailored losses, ensuring matting-specificity and subtle modifications on the portrait. We demonstrate that the proposed method can refine real portrait images for arbitrary matting models, boosting the performance of automatic alpha matting by a large margin. In addition, we leverage the generative property of StyleGAN, and propose to generate enhanced portrait data which can be treated as the pseudo GT. It addresses the problem of expensive alpha matte annotation, further augmenting the matting performance of existing models.

Index Terms—alpha matting, latent space, generative model

I. INTRODUCTION

Self-portrait photographs, i.e., selfies, have been popular on social media, and its related portrait enhancement applications are widely developed to fulfill different aesthetic requirements. A fundamental technique that drives these applications is image matting, which is to estimate the foreground transparency of portrait images. Image matting assumes an image is composed of the linear combination of the foreground and background with the alpha matte \( \alpha \) indicating the foreground transparency. It can be formulated as

\[
I_i = \alpha_i F_i + (1 - \alpha_i) B_i, \quad \alpha_i \in [0, 1],
\]

where \( i = (x, y) \) denotes the pixel position in the image.

It can be seen that image matting is an ill-posed problem given only 3 known values but 7 values need to be solved for each pixel. Therefore, most matting methods \cite{28}, \cite{3}, \cite{52}, \cite{30}, \cite{48}, \cite{7} introduce an additional user-specified scribble or trimap to reduce the dimension of the solution space. This kind of auxiliary input is a confidence map that gives each pixel of the image a probability of belonging to the foreground, background or a mixed area. These works have been proven to achieve a reliable result. However, they highly rely on a well-defined trimap, and this precise prior knowledge is always hard to obtain. Their results deteriorate rapidly if the trimap is defective, which is a practical problem in real-world scenarios. For this reason, automatic portrait matting has been proposed in recent years \cite{40}, \cite{48}, \cite{36}, \cite{7}.

Notwithstanding the demonstrated success, automatic portrait matting still suffers from ambiguous semantic and disreputable alpha values. These problems can be hardly addressed without prior knowledge due to the entanglement between foreground and background colors. However, we observe that foreground and background may be disentangled if small changes in the image are introduced (see Fig. 1a, 1b, and 1c), as the alpha matte is estimated from a global perspective. This finding is inspired by sampling-based methods \cite{18}, \cite{42}, as changing the foreground or background slightly may provide a completely different set of support samples for alpha value estimation.

Based on the above observation, we propose an alternative

![Fig. 1: We propose to generate matting-specific portraits by exploring the latent space of GANs. (a), (b) show that subtle changes in the portrait can significantly enhance the output alpha matte (see blue boxes and red arrows). (c) shows the learned noise perturbations (rescaled and remapped) are automatically attached in high-frequency image contents so that they are not sensitive to human perception. (d) shows that the proposed method can be also used to synthesize pseudo GT that augments existing matting models.](image-url)
solution for resolving the ill-posed matting problem. Unlike traditional image matting that devoted to alpha matte prediction, we focus on how an input image affects the output alpha matte. However, connecting image editing with alpha matting is not trivial. Naively cascading two modules lacks interpretable control of image editing, and therefore cannot guarantee subtle changes applied in the portrait. In addition, this requires a conditional Generative Adversarial Network (GAN) structure which cannot reconstruct high-resolution and high-fidelity portrait.

In this paper, we achieve matting-specific portrait enhancement by exploring the latent space of GAN models. Very recent literature [41], [37], [44] show that meaningful directions can be discovered in the latent space of GAN models that correspond to human interpretable image transformation. Except for editing randomy generated images, latest works [1], [2], [17], [55] propose to invert a real image back to the latent code of the GAN model for reconstruction. All these prior works motivate us to explore the following question: is there an enhanced portrait that exists in the latent space that is more compatible with a reference matting model? However, the main barrier is that there is no labeled data for alpha matting to regress to during testing. Our principle is to convert the regression problem to a classification problem so that it can be optimized in a self-supervised manner. To this end, we first decompose the intractable automatic portrait matting problem into two parts, ternary trimap prediction and alpha matte estimation. The former serves as an optimizer while the latter serves as an assessor. In this way, we can optimize the latent code governed by a reference matting model with four tailored losses, and our objective is to discover a new latent code that minimizes the prediction entropy and corrects the generative distribution in a self-supervised manner. To ensure matting-specificity as well as subtle modifications on the portrait, we search the optimal latent code in the noise latent space $\mathcal{W}$ and the intermediate latent space $\mathcal{W}^+$ of StyleGAN.

The input latent code can be either inverted from a real image or randomly-generated, each of which enables two different applications (see Fig. 1). The former one is used to enhance a real portrait for a better matting performance, while the latter can create a massive amount of pseudo GT data for augmenting the performance of arbitrary matting models. Extensive experiments demonstrate that the proposed two alternative solutions can enhance a portrait with minimum modifications or produce matting-specific portrait data, boosting state-of-the-art models by a large margin.

Our contributions are summarized as follows:

- Different from existing image matting works that predict an alpha matte based on the fixed input image, we focus on how an input image affects the output alpha matte. We seek an enhanced image by exploring the latent space of GAN that is more compatible with a referenced matting model in a self-supervised manner.

- We present a portrait enhancement approach that can refine an input portrait with minimum changes in the image. State-of-the-art models can be boosted by a large margin on two datasets.

- We propose to generate matting-specific portrait data, addressing the expensive annotation problem of alpha matte. The synthesized data can augment state-of-the-art matting models.

II. RELATED WORK

A. Natural Image Matting

Traditional image matting can be categorized into sampling-based and affinity-based approaches. Sampling-based approaches [18], [21], [9], [45] assume that the alpha values for two neighboring pixels are similar if their colors are similar. Affinity-based approaches [3], [8], [27] propagate alpha values from the known regions into unknown ones by the affinities of neighboring pixels. Several deep learning-based approaches are proposed for image matting to leverage the powerful representation learning ability. Xu et al. [48] propose a large scale dataset and the first attempt to apply a neural network for nature image matting. Li et al. [30] present the Guided Contextual Attention to propagate high-level opacity information based on the learned low-level affinity. Lu et al. [35] develop an IndexNet to generate the indices on the feature map adaptively. SampleNet [42] estimates the background and foreground colors of pixels in unknown regions using deep neural networks prior. Cai et al. [5] aim to refine the trimap by an adaptation module in the matting network. However, these works rely heavily on the user-designed trimap. When the trimap is unavailable, a typical solution is to predict the trimap by a semantic segmentation model [40], [7]. However, compared with the user-designed trimap, the predicted trimap is insufficiently accurate. Background matting [39], [31], [10] can extract alpha mattes without the need of trimap, but it requires an additional input of clean background image, which confines the flexibility. Unlike previous approaches, we propose an alternative solution for automatic portrait matting that focuses on how an input image affects the output alpha matte by exploring the latent space of GAN models.

B. Image Editing via Latent Space Exploration

Generative models show great potential in synthesizing various images by taking random latent codes as inputs. It’s evidenced that a pre-trained generator encodes rich knowledge prior, which benefits many downstream works, such as face swapping, super-resolution, semantic editing, and so on [53], [49], [54], [41], [17]. Particular, there are many image editing methods explore the latent space of a pre-trained generative model (e.g., StyleGAN) in supervised or unsupervised manners [41], [17], [55], [14], [26], [51]. Different from existing works that learn the end-to-end models for image enhancement or restoration [15], [29], [11], GAN-based editing methods need to invert a target image to the latent code in advance, which is called GAN inversion [1], [2], [50], [46], [43]. GAN inversion works discover a latent code that can produce the closest result to the input real image. In particular, Abdal et al. [1] propose an optimization-based algorithm that optimizes the latent code in the $\mathcal{W}^+$ of a fixed StyleGAN governed by pixel-wise reconstruction. Their subsequent work [2] introduces noise optimization as a complement to the $\mathcal{W}^+$ latent space for restoring the high-frequency details in images. Zhu et al.
al. [55] propose to learn an encoder that ensures the inverted code in the semantic domain of the original latent space and explores many semantic image editing tasks. Compared to these works, we have a different objective that aiming at producing matting-specific portrait. We emphasize editing portraits with the minimum changes while making them the most “predictable” by a matting model.

III. METHODOLOGY

A. Problem Formulation

Given an input portrait $I_o$, a reference matting model $M_o$, and a pre-trained trimap model $M_t$, our goal is to obtain an enhanced portrait produced by the generator $G$ of StyleGAN [22], [23]. Unlike traditional GANs, the output of StyleGAN is controlled by two latent codes from $Z$ and $N$ spaces. As shown in Fig. 2, StyleGAN maps the latent code from $Z$ space to intermediate latent space $W$ space with few fully-connected layers for a better semantic disentanglement [22], [23]. Abdal et al. [1] further extend the $W$ space to $W^+$ space for a better representative ability, and many GAN inversion works invert images in this space [50], [46], [43]. Besides, noise latent space $N$ provides high-frequency information to the generator. It produces the noise latent code to each scale of the synthesis network. As shown in [22], [2], this space controls subtle details of the synthesis.

We take two latent vectors $(w_o, n_o)$ in $W^+$ and $N$ spaces to represent the input portrait $I_o$, and these vectors can be the inverted codes using GAN inversion methods [1], [2] or randomly generated ones. Our objective is to obtain the optimal latent vectors $(w^*, n^*)$ as follows

$$ (w^*, n^*) = \arg \min_{(w,n)} \lambda_1 L_{em} + \lambda_2 L_{ca} + \lambda_3 L_{pc} + \lambda_4 L_{pp}, $$

where $L_{em}$, $L_{ca}$, $L_{pc}$, $L_{pp}$ denote the entropy minimization loss, compositional adversarial loss, portrait consistency loss, and perceptual loss respectively. $\lambda_x$ is the balancing weight. These losses constrain the searching of the enhanced portrait $G(w^*, n^*)$ to be matting-specific and face-preserved. Fig. 3 shows the pipeline of our optimization.

1) Entropy Minimization Loss: The first goal of our portrait enhancement is to alleviate the difficulty of separating the foreground and background. As there is no ground-truth for training the regression problem of matting, we instead propose to minimize the prediction entropy. The rationale behind this loss is to edit the portrait so that the matting model predicts a high-confidence alpha matte, i.e., refining ambiguous regions of the alpha matte. However, alpha matte estimation is a regression problem without explicit confidences during inference. To resolve this problem, we shift our optimization target to a classification problem, which indeed equals to the essential trimap prediction (i.e., classifying foreground, background, and unknown regions) in the pipeline. The new optimization task shares the same goal of obtaining the enhanced portrait, as the quality of a trimap is proportional to the quality of the predicted alpha matte. Following the entropy minimization principle [34], [33], [16], [47], we encourage the low-density separation between classes by minimizing the entropy of confidence map, which can be represented as

$$ L_{em} = H(M_t(G(w, n))), $$

where $H(x) = -\sum x \log(x)$ is the entropy function, and $M_t(G(w, n))$ represents the confidence map of $G(w, n)$ predicted by the trimap model $M_t$. In this way, we refine the portrait $G(w, n)$ so that a more certain confidence map can be obtained.

2) Compositional Adversarial Loss: As the entropy minimization loss is not directly applied to the alpha matte, it is intractable to evaluate whether the enhanced alpha matte works better than the original one, especially without ground-truth data.

We further introduce a compositional adversarial loss to evaluate the alpha matte. Instead of directly examining the alpha matte, we use the enhanced alpha matte, with a new background image randomly selected from MS COCO dataset [32] to compose a new portrait image for evaluation. The proposed compositional adversarial loss aligns the distance between distributions of real portraits and the composite ones, yielding a better alpha matte. Specifically, we utilize a pre-trained discriminator $D$ of StyleGAN to distinguish the composite portrait as a fake sample, which is defined as follows

$$ L_{ca} = -\log(D(C(F, B_r, \alpha_e))), $$

where $C(\cdot)$ is a composition function that composes a new portrait with the enhanced alpha matte $\alpha_e$, foreground portrait $F$, and a random background image $B_r$.

We use Closed-form Matting [27] for estimating the foreground portrait $F$ based on the enhanced matte $\alpha_e$. Particularly, given an alpha matte $\alpha$, Closed-form Matting [27] estimates the foreground $F$ and background $B$ by minimizing a cost function for each pixel $i$ and color channel $c$: Fig. 2: Illustration of different latent spaces in StyleGAN.
Perceptual loss is defined as pixel-wise practical importance. Therefore, we propose to minimize the loss.

$$
\text{cost}_{\text{global}}(F, B) = \sum_{i \in I} \sum_{c}(\alpha_i F_i^c + (1 - \alpha_i) B_i^c - I_i^c)^2 + |\alpha_i| \left( (F_i^c)^2 + (B_i^c)^2 \right) + |\alpha_i| \left( (F_i^o)^2 + (B_i^o)^2 \right).
$$

The first term aligns the distance between the composed and original images, and the rest two terms reduce the magnitude of color gradients $F_i^c$, $F_i^o$, $B_i^c$, and $B_i^o$, in regions of large $\alpha$-gradients $|\alpha_i|$ and $|\alpha_i|$ for preserving the texture information.

3) Portrait Consistency Loss: In the scenario of using a real portrait, maintaining the original appearance is of great practical importance. Therefore, we propose to minimize the pixel-wise $L_2$ distance between the enhanced portrait image $G(w^*, n^*)$ and the original one $I_o$ to constrain the editing degree. Portrait consistency loss is defined as follows

$$
\mathcal{L}_{pc} = \|G(w^*, n^*) - I_o\|_2.
$$

4) Perceptual Loss: Besides the portrait consistency loss, we also utilize a perceptual loss [20] for maintaining the subtle facial details between two portraits. We compute the $L_2$ distance between features produced by the enhanced portrait image $G(w^*, n^*)$ and the original one $I_o$ after feeding to a pre-trained VGG-16 network. Perceptual loss is defined as

$$
\mathcal{L}_{pp} = ||\Phi_{VGG}(G(w^*, n^*)) - \Phi_{VGG}(I_o)||_2,
$$

where $\Phi_{VGG}()$ is a pre-trained VGG-16 network and we select the features produced by conv4_2 layer for calculating the loss.

**Algorithm 1** Optimizing the latent codes for the enhanced portrait and alpha matte

**Input:** Original portrait image $I_o$, random background image $I_B$, pre-trained trimap model $M_t$, pre-trained matting model $M_a$, pre-trained StyleGAN $G$ and $D$, iterative times $T$.

**Output:** Enhanced portrait $G(w^*, n^*)$ and alpha matte $\alpha_e$.

1: begin  
2: Invert the portrait image $I_o$ to the latent vectors $(w_o, n_o)$ under the supervision of $\mathcal{L}_{pc}$ and $\mathcal{L}_{pp}$, and $(w, n) = (w_o, n_o)$.  
3: for steps = 1, 2, ..., $T$ do  
4: Generate an enhanced portrait image $G(w, n)$ by $(w, n)$ to $G$.  
5: Predict the trimap of $G(w, n)$ by $M_t$.  
6: Estimate the enhanced alpha matte $\alpha_e$ by $M_a$.  
7: Estimate the foreground layer $F$ using the enhanced alpha matte $\alpha_e$ by minimizing cost function $\text{cost}_{\text{global}}(F, B)$.  
8: Compose a new portrait with a random background image $B_r$, foreground image $F$, and $\alpha_e$ based on Eq. (1).  
9: Optimize the latent codes $(w, n)$ with Eq. (2).  
10: end for  
11: $(w^*, n^*) = (w, n)$.  
12: return $G(w^*, n^*)$ and $\alpha_e$.  
13: end begin

**B. Portrait Enhancement**

Although the proposed optimization is performed on the latent space of a generative model, it can be applied for real image enhancement. Given a portrait image $I_o$, we apply GAN
weights in Eq. (2) as to optimize the latent codes by Adam optimizer [25]. We use 1024 codes again to obtain the enhanced portrait \( G \) space \( N \) consists of 2,000 samples with a resolution of 800 × 800, we resize the original portraits \( \lambda=1 \). Note that the portrait consistency loss \( \mathcal{L}_{pp} \) and perceptual loss \( \mathcal{L}_{pp} \) are removed as no original image is needed, and it can also relax the optimization with a large searching space. In this way, we can generate arbitrary numbers of matting-specific portraits. All the predicted alpha mattes are of high-confidence that can be treated as pseudo GT data. Note here we use the initial generated portraits, together with the enhanced alpha mattes to form the pseudo GT pairs. Here we use the initial portraits instead of the enhanced portraits, as the enhanced portraits are likely to be overfitted to the pre-trained model, and provide less supervision information.

C. Portrait Generation

An important property of using GANs is the flexibility of generating data. We first leverage StyleGAN to generate a large number of portrait images with random latent vectors. Then these images are enhanced by discovering the optimal latent codes using Eq. (2). Note that the portrait consistency loss \( \mathcal{L}_{pc} \) and perceptual loss \( \mathcal{L}_{pp} \) are removed as no original image is needed, and it can also relax the optimization with a large searching space. In this way, we can generate arbitrary numbers of matting-specific portraits. All the predicted alpha mattes are of high-confidence that can be treated as pseudo GT data. Note here we use the initial generated portraits, together with the enhanced alpha mattes to form the pseudo GT pairs. Here we use the initial portraits instead of the enhanced portraits, as the enhanced portraits are likely to be overfitted to the pre-trained model, and provide less supervision information.

IV. Experiments

A. Implementation Details

We implement the proposed method in Pytorch on a PC with an Nvidia GeForce RTX 2080Ti. The proposed method takes about 3 mins to optimize the latent codes for each image. We utilize the generator and discriminator of StyleGAN2 [23] pre-trained on the FFHQ dataset [22]. Since the output resolution of StyleGAN2 is 1024 × 1024, we resize the original portraits to the resolution of 1024 × 1024 for portrait enhancement. We optimize the latent codes by Adam optimizer [25]. We use 500 gradient descent steps with a learning rate of 0.0001 and a weight decay of 0.0005. We empirically set the balancing weights in Eq. (2) as \( \lambda_1 = 1 \), \( \lambda_2 = 1 \), \( \lambda_3 = 10 \), and \( \lambda_4 = 10 \).

B. Dataset and Evaluation Metrics

We conduct our experiments on two datasets, the first one is the largest portrait matting dataset [40]. It was collected from Flickr with a good variety of age, color, clothing, and hairstyle. Alpha mattes are annotated with intensive user interactions. It consists of 2,000 samples with a resolution of 600 × 800, and we follow [40] to divide images into training and testing sets with 1,700 and 300 images, respectively.

Although the above dataset is the largest portrait matting dataset, it is not fully labeled by human. To further evaluate our method on high-quality data, we select portrait images from the test set of Adobe Image Matting Dataset [48], resulting a new dataset that consists of 140 images.

We use four quantitative metrics to evaluate the alpha matting results, namely, the mean square error (MSE), the sum of absolute differences (SAD), the gradient error (Grad), and connectivity errors (Conn) [38]. MSE and SAD mainly focus on the statistical differences between two alpha mattes, but they may not correlate to the visual quality perceived by human [38]. Compared with MSE and SAD, Conn and Grad are more related to human perception. Specifically, Conn evaluates the disconnected foreground objects and Grad mainly concentrates on the over-smoothed or erroneous discontinuities in the alpha matte.

C. Trimap and Matting Models

DeepLabV3 [6] with a backbone of ResNet50 [19] pre-trained on the CoCo-Stuff Dataset [4] is employed as our trimap model. Following the work [5], we also use alpha mattes as the ground-truth for fine-tuning the trimap model. We take six state-of-the-art matting methods, Deep Image Matting (DIM) [48], IndexNet [35], Guided Contextual Attention (GCA) [30], AdaMatting [5], SampleNet [42] and MODNet [24] as the reference matting models. Note that all methods need trimap as the extra input except MODNet [24]. MODNet is a trimap-free matting method, and only the loss of \( L_{ca} \) can be applied to this model. All of them are pre-trained on the Adobe Image Matting Dataset [48] and then fine-tuned on the portrait matting dataset [40] except SampleNet [42]. Because it requires two inputs of foreground and background images, therefore we cannot fine-tune their model using the portrait matting dataset [40] or our generated portraits. Instead, we directly use the model pre-trained on Adobe Image Matting Dataset [48] for enhancement, which also demonstrates the generalization ability of our method.

D. Evaluation on Portrait Enhancement

In this section, we evaluate the effectiveness of the proposed portrait enhancement method. The performances of matting models on the original portraits are set as baselines. Note that we froze all parameters of the matting model and only optimize the latent space to find the optimal high-dimensional representation that most compatible with these matting models. We show both quantitative and qualitative results in the following subsections.

1) Quantitative Results: Quantitative evaluation on the proposed portrait enhancement approach is shown in Tab. 1. A clear performance boosting can be found for all the methods on both two datasets. Particularly, perceptual-based metrics Grad and Conn reflect the enhancement property of our method. Both these two metrics focus on evaluating continuity of the generated alpha mattes, which is the main source affects the human perception of the new composite images. As can be seen, the proposed enhancement improves existing methods by a large margin for all trimap-based matting models, e.g., an average Grad improvement of 11.32% on portrait dataset and 20.732% on adobe matting dataset has been obtained. This can be specifically attributed to our entropy minimization strategy and compositional loss, as discontinuous regions are mostly with a low-degree of confidence and can be easily detected in the composite images. As a result, the proposed enhancement mostly corrects discontinuity (as shown in Sec. IV-D2) and more favorable to human perception. In terms of error measurement, both the MSE and SAD achieves at least 4%
improvement for all the six methods on the portrait dataset. In particular, although the room for MSE improvement is small, the best two methods DIM and IndexNet increase by more than 10%. SAD metric shows a larger performance gain in number (0.933 improvement for IndexNet). These statistical improvements demonstrate that the proposed enhancement approach removes prediction outliers very well. Given the predicted trimap in the trimap-free method MODNet [24], our method can also improve its performance on all metrics. Overall, the proposed enhancement shows great improvements on different metrics (up to 10% for most metrics), and also demonstrates the generality for state-of-the-art methods (most methods are improved by a large margin).

2) Qualitative Results: Fig. 4 shows the qualitative results of portrait enhancement. As the proposed entropy minimization loss is directly applied to trimap prediction, we can see from the second column of each group, the enhanced trimaps fixed discontinuous regions, leading to better alpha mattes and composition results (see blue boxes). On the other hand, as can be seen in the perturbation images, the enhanced portraits add subtle details in almost all the areas. This demonstrates our observation that minor changes can yield a significantly different alpha matte. An interesting observation is that noise patterns are mostly attached to high-frequency components of the images, e.g., edges. This automatically learned operation surprisingly agrees with human perception, as humans are not particularly sensitive to high-frequency content in images. We further observe that relatively intense noise patterns may add to background regions, as shown in the top-left example in Fig. 4. This interesting observation shows that the network is “taught” to avoid adding obvious noise to the foreground. We believe this attributes to the proposed compositional adversarial loss, as it ensures the high quality of composite images. These meaningful patterns are the reason of correcting those discontinuous holes in the alpha mattes.

E. Evaluation on Portrait Generation

Here we evaluate the performance of our second application, portrait generation. We use four state-of-the-art methods to govern the optimization, leading to four different sets of synthesized data. There might be a problem that the generated portrait-matte pairs overfit to a reference matting model and gain less improvement. Hence we mix the synthesized samples that come from different matting models for fine-tuning a matting model. This strategy is to increase the diversity of the data, as different models have different emphasizes on easy-to-extract images. Besides, it also avoids the overfitting problem, as image pairs generated by different matting models are blind to the target one. Each set contains 10,000 synthesized images. These four sets of data, combined with the original training data of [40], are used to fine-tune four matting models.

Tab. II shows the quantitative evaluation on portrait augmentation with two datasets. We can see that all the state-of-the-art models improve their performances to a certain extent. This implies that the synthesized data is of high-confidence GT, effective for training matting models. Compared with portrait enhancement, fine-tuning with synthesized data results in less improvement. This is because although the network is fine-tuned with enhanced portraits, applying enhancement in the testing phrase optimizes an image-specific latent code that leads to the best matting performance.

We also show some examples of the generated portraits and the corresponding alpha mattes in Fig. 5. As can be seen, the generated portraits are diverse enough, with high-quality alpha mattes without discontinuous regions.

F. Ablation Study

In this section, we analyze the efficacy of the optimization in different latent spaces and loss functions, under the application of portrait enhancement on Portrait Matting Dataset. We first compare with six variants: 1) optimizing the latent code \(w\) only while keeping the noise code \(n\) fixed (\(w^*, n\)); 2) optimizing the noise code \(n\) only (\(w, n^*\)); 3) without the entropy

### Table I: Quantitative evaluation on portrait enhancement with respect to six matting models on two datasets. The lower the better for all metrics. Top 2 improved methods (in percentage) for each metric are marked in blue and green respectively.

| Models      | Metrics   | Grad↓ | Conn↓ | MSE↓ | SAD↓ |
|-------------|-----------|-------|-------|------|------|
|             | Ori.      | Enh.  | ∆ / [%Δ] | Ori. | Enh. | ∆ / [%Δ] | Ori. | Enh. | ∆ / [%Δ] | Ori. | Enh. | ∆ / [%Δ] |
| Portrait Matting Dataset [40] |            |       |       |      |      |       |      |      |       |      |      |       |
| DIM [48]    | 6.895     | 7.459 | -1.240/14.25% | 12.291 | 11.374 | -0.9177/46% | 0.017 | 0.015 | -0.002/11.76% | 13.269 | 12.464 | -0.805/6.07% |
| IndexNet [35] | 7.405     | 6.168 | -1.237/16.70% | 11.345 | 10.302 | -1.043/19% | 0.016 | 0.014 | -0.002/12.5% | 11.826 | 10.799 | -1.027/6.88% |
| GCA [30]    | 9.170     | 8.115 | -1.055/11.50% | 11.382 | 10.877 | -0.505/43% | 0.017 | 0.016 | -0.001/5.88% | 11.751 | 11.384 | -0.367/12% |
| AdaMatting [5] | 21.882    | 21.186 | -0.696/31.9% | 18.291 | 17.458 | -0.833/45.5% | 0.029 | 0.028 | -0.001/3.44% | 18.225 | 17.664 | -0.561/0.8% |
| SampleNet [42] | 8.511     | 7.703 | -0.808/4.9% | 11.119 | 10.132 | -0.9878/87% | 0.017 | 0.016 | -0.001/5.88% | 11.503 | 10.686 | -0.8357/26% |
| MODNet [24] | 22.436    | 21.552 | -0.884/3.94% | 17.523 | 17.227 | -0.296/1.69% | 0.028 | 0.027 | -0.001/3.57% | 19.359 | 18.907 | -0.452/2.33% |
| Adobe Image Matting Dataset [48] |            |       |       |      |      |       |      |      |       |      |      |       |
| DIM [48]    | 21.718    | 15.589 | -6.129/28.22% | 23.364 | 21.366 | 1.99/8.55% | 0.038 | 0.037 | -0.002/2.63% | 23.417 | 22.766 | -0.651/7.8% |
| IndexNet [35] | 21.268    | 18.617 | -2.651/12.46% | 22.541 | 21.144 | -1.3976/20% | 0.038 | 0.037 | -0.002/2.63% | 22.701 | 22.323 | -0.378/1.67% |
| GCA [30]    | 22.489    | 17.261 | -5.226/23.25% | 22.106 | 21.165 | -0.941/24.6% | 0.038 | 0.037 | -0.002/2.63% | 22.205 | 21.445 | -0.76/4.42% |
| AdaMatting [5] | 24.934    | 19.577 | -5.357/23.35% | 23.274 | 21.953 | -1.981/51% | 0.041 | 0.039 | -0.002/4.88% | 25.613 | 24.655 | -0.958/0.374% |
| SampleNet [42] | 21.806    | 18.729 | -3.077/14.11% | 23.029 | 21.527 | -1.502/6.52% | 0.038 | 0.037 | -0.002/2.63% | 22.345 | 22.003 | -0.342/1.53% |
| MODNet [24] | 25.783    | 25.042 | -0.741/2.96% | 24.433 | 23.903 | -0.53/2.17% | 0.045 | 0.042 | -0.003/6.67% | 27.335 | 27.029 | -0.306/1.11% |
Fig. 4: Qualitative evaluation on portrait enhancement. The odd row images are original portraits with their corresponding predicted trimap, alpha matte, perturbations (rescaled and remapped), and the composite image (with a white background). The even row images show the enhanced results and ground truth (d).

minimization loss (w/o $L_{cm}$); 4) without the compositional adversarial loss (w/o $L_{ca}$); 5) without the portrait consistency loss (w/o $L_{pc}$); 6) without the perceptual loss (w/o $L_{pp}$). For variants 1) and 2), we optimize them using all the four losses. Similarly, for variants 3), 4), 5) and 6), we optimize the latent codes $w$ and $n$ simultaneously. In addition, we also evaluate out method with various hyper-parameter in Eq. 2, and this results in two variants: 7) Set $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 1$, and $\lambda_4 = 1$ in Eq. 2 (Hyper-param. #1); 8) Set $\lambda_1 = 5$, $\lambda_2 = 5$, $\lambda_3 = 10$, and $\lambda_4 = 10$ in Eq. 2 (Hyper-param. #2). Furthermore, for evaluating the influence of different foreground estimation methods, we use MLFgE [13] and Blur-Fusion [12] instead of Closed-form Matting [27] for the foreground estimation respectively, resulting in two variants: 9) FG w/ [13] and 10) FG w/ [12]. 11) We also use the GT matte instead of enhanced matte for the foreground estimation used in Eq. 5, and name this variant as (FG w/ GT $\alpha$).

We take IndexNet as the reference matting model (also
TABLE II: Quantitative evaluation on portrait augmentation with respect to five matting models on two datasets. “FT” represents the results of fine-tuned model. The lower the better for all metrics. Top 2 improved methods (in percentage) for each metric are marked in blue and green respectively.

| Models          | Metrics | Grad↓ | Conn↓ | MSE↓ | SAD↓ |
|-----------------|---------|-------|-------|------|------|
|                 |         | Ori   | FT    | Δ / [%Δ] | Ori   | FT    | Δ / [%Δ] | Ori   | FT    | Δ / [%Δ] | Ori   | FT    | Δ / [%Δ] |
| DIM [38]        |         | 8.699 | 8.212 | -0.487/5.60% | 12.291 | 11.721 | -0.570/4.64% | 0.017 | 0.016 | -0.001/5.88% | 13.269 | 12.973 | -0.296/2.23% |
| IndexNet [35]   |         | 7.405 | 7.029 | -0.376/5.08% | 11.345 | 10.933 | -0.412/3.63% | 0.016 | 0.015 | 0.001/6.25% | 11.826 | 11.472 | -0.354/2.99% |
| GCA [30]        |         | 9.170 | 8.913 | -0.257/2.80% | 11.382 | 11.131 | -0.251/2.21% | 0.017 | 0.016 | -0.001/5.88% | 11.751 | 11.428 | -0.323/2.75% |
| AdaMatting [5]  |         | 21.882 | 21.037 | -0.845/3.86% | 18.291 | 17.776 | -0.515/2.82% | 0.029 | 0.027 | 0.002/6.90% | 18.225 | 17.854 | -0.371/2.04% |
| MODNet [24]     |         | 22.436 | 22.033 | -0.403/1.79% | 17.523 | 17.374 | -0.490/8.5%  | 0.026 | 0.027 | -0.001/3.57% | 19.359 | 18.871 | -0.488/2.52% |

| Portrait Matting Dataset [40] |         | 20.224 | 7.404 | -0.001/2.63% | 10.657 | 10.302 | -0.371/2.04% | 0.015 | 0.016 | -0.001/5.88% | 17.374 | 16.921 | -0.323/2.75% |
| Adobe Image Matting Dataset [48] |         | 18.871 | 18.871 | -0.251/2.21% | 10.799 | 10.799 | -0.371/2.04% | 0.015 | 0.016 | -0.001/5.88% | 18.225 | 17.854 | -0.371/2.04% |

**Fig. 5:** Examples of the generated portraits and their corresponding alpha mattes. The synthesized portraits are diverse and no discontinuous region appears in the pseudo GT data.

TABLE III: Comparisons with various variants using IndexNet as the reference matting model, under the application of portrait enhancement. “Baseline” represents the performance on the original portraits and “(w, n)” represents the performance of our complete model. The lower the better for all metrics and the best results are marked in bold.

| Metrics | Variants | Baseline | (w, n) | (w, n*) | w/o $L_{cm}$ | w/o $L_{ca}$ | w/o $L_{pc}$ | w/o $L_{pp}$ | (w, n*) |
|---------|----------|----------|--------|---------|--------------|--------------|--------------|--------------|---------|
| Grad↓   |          | 7.404    | 6.751  | 6.688   | 7.015        | 6.368        | 6.613        | 6.323        | 6.168   |
| Conn↓   |          | 11.345   | 10.657 | 11.013  | 11.082       | 10.528       | 10.688       | 10.378       | 10.302  |
| MSE↓    |          | 0.016    | 0.015  | 0.015   | 0.015        | 0.014        | 0.014        | 0.014        | 0.014   |
| SAD↓    |          | 11.824   | 11.154 | 11.477  | 11.568       | 10.994       | 10.893       | 10.956       | 10.799  |

served as the baseline). The quantitative comparisons are shown in Tab. III and Tab. IV. We can see that every variant outperforms the baseline on all metrics. This indicates the proposed idea is effective, even is conducted in one latent space with incomplete losses. Meanwhile, the performances of variants (w, n) and (w, n*) show that the latent code w contributes more than code n to enhancement, and optimizing both latent codes w and n significantly boost the performance.

Regarding the loss functions, it is unsurprising that the proposed entropy minimization loss ($L_{cm}$) contributes the most to the improvement. This is because $L_{cm}$ is the core component of the latent codes searching process that modifies uncertain and highly-difficult regions. Meanwhile, we observe that deactivating portrait consistency loss (w/o $L_{pp}$) and perceptual loss (w/o $L_{pc}$) decrease the final performance slightly. Although these two losses are introduced to maintain the fidelity of the original portrait, the input GAN inverted latent codes are able to generate similar portraits. As a result, using one of these losses is sufficient to ensure the fidelity of the original input. Moreover, both variants (Hyper-param. #1) and (Hyper-param. #2) present similar performances with the complete model (w, n*), which shows our model is not sensitive to hyper-parameters in Eq. 2. For the variants (FG w/ [13]) and (FG w/ [12]), they report similar results to the complete model (w, n*), which indicates that our method is not sensitive to the foreground estimation methods. The variant (FG w/ GT α) presents a similar performance with foreground estimation using the enhanced matte (w, n*). That shows
TABLE IV: Comparisons with various variants with different parameter settings and foreground estimation methods. The lower the better for all metrics and the best results are marked in bold.

| Metrics | Variants | Hyper-param. #1 | Hyper-param. #2 | FG w/ [13] | FG w/ [12] | FG w/ GT α | (w*, n*) |
|---------|----------|----------------|----------------|-----------|-----------|------------|-------|
| Grad↓   |          | 6.165          | 6.166          | 6.165     | 6.164     | 6.165      | 6.168 |
| Conn↓   |          | 10.293         | 10.312         | 10.292    | 10.303    | 10.296     | 10.302 |
| MSE↓    |          | 0.014          | 0.014          | 0.014     | 0.015     | 0.014      | 0.014 |
| SAD↓    |          | 10.899         | 10.896         | 10.895    | 10.901    | 10.893     | 10.799 |

Fig. 6: Our method can handle challenging cases like profiles (left part). In addition, we can also enhance the portraits with low-quality initial trimaps or alpha mattes (right part).

our method is not restricted by the quality of the foreground layer. It is because 1) compared with the entropy minimization loss, the compositional adversarial loss contributes less to the improvement, and 2) the quality of the composited sample is mainly influenced by the matting quality. For example, composing a GT foreground with a discontinuous alpha matte (like holes) will produce a bad composition quality. Therefore, rather than noticing the insignificant differences in the foreground estimation, the discriminator can easily classify the discontinuity and provides feedback for training a better alpha matte.

G. Evaluation on Challenging Samples

In this section, we evaluate our method on challenging samples. We first give the evaluation on profile faces. As shown in the left part of Fig. 6, our method also fixes the undesired holes on the trimap, or correct the misclassified background (see the sample at the bottom-left in Fig. 6). This reveals that the proposed method is not limited to the frontal faces but can handle profiles as well.

We also present the improvement on those samples that have poor initial performances in the right part of Fig. 6. For example, the portrait at the bottom-right in Fig. 6 contains complicated background, and the original trimap has a large erosion around the hat region. Meanwhile, the original matte also presents discontinuous regions. For those low initial quality cases, we can also remove the erosion on the trimap and improves the quality of matte after optimization, which shows that our method can boost the performance of any input portrait samples.

V. Conclusion, Limitation, and Future Work

In this paper, we propose an alternative solution for image matting. We demonstrate that small changes in the input image may significantly enhance the matting performance. These small changes are discovered by optimizing the latent vectors in the latent spaces of GAN models. We tailor four losses to obtain the optimal latent vectors, easing the matting difficulty with minor modifications in the portrait. The proposed method enables two applications, portrait enhancement and generation. The first one can enhance a real portrait such that the performance of existing matting models can be largely improved. The second application generates high-confidence pseudo data, augmenting the representation ability of existing matting models.

The limitation of our method is that we need to optimize the latent code for each image. This process takes around 3 mins and it might limit our practical usage. Another limitation is
that combining both the portrait enhancement and generation cannot yield better performance, and the portrait enhancement still has the best result. This is because the synthesized data is also the enhanced portraits, training with these images cannot beat the optimized latent code using portrait enhancement. In the future, we aim to discover whether there is a general direction exists in the latent space, such that moving along that direction can produce a matting-specific version.

REFERENCES

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In ICCV, pages 4432–4441, 2021.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In CVPR, pages 8296–8305, 2020.

[3] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In CVPR, pages 29–37, 2017.

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In CVPR, pages 1209–1218, 2018.

[5] Runhao Gai, Xiaoxiao Zhang, Haoqiang Fan, Haibing Huang, Jiangyu Liu, Jianming Liu, Jiaqing Liu, Jue Wang, and Jian Sun. Disentangled image matting. In ICCV, pages 8819–8828, 2019.

[6] 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. In CVPR, pages 770–778, 2016.

[7] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu. Modnet: real-time trimap-free portrait matting via objective decomposition. In AAAI, 2020.

[8] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. In CVPR, pages 4432–4441, 2019.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

[10] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In NeurIPS, pages 5967–5976, 2017.

[11] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. IEEE TPAMI, 30(2):228–247, 2007.

[12] A. Levin, A. Rav-Acha, and D. Lischinski. Spectral matting. IEEE TPAMI, 30(10):1699–1712, 2008.

[13] Boyun Li, Yuanhao Gao, Shuhang Gu, Jerry Zitao Liu, Jiey Tianyi Zhou, and Xi Peng. You only look yourself: Unsupervised and untrained single image dehazing neural network. IJCV, 129(5):1754–1767, 2021.

[14] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. High-fidelity gan inversion for image attribute editing. In NeurIPS, volume 34, pages 11450–11457, 2020.

[15] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the code gan prior. In CVPR, pages 4432–4441, 2020.

[16] Shanchuan Lin, Andrew Ryabtsev, Soumyadip Sengupta, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In CVPR, pages 8762–8771, 2021.

[17] Carlos Orrite, Miguel Angle Varona, Eduardo Estopiñán, and José Ramón Beltrán. Portrait segmentation by deep refinement of image matting. In ICIP, pages 1495–1499, 2019.

[18] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelet. Controlling generative models with continuous factors of variations. In ICLR, 2020.

[19] Xiaoyang Shen, Jinyu Shen, Jinjin Gu, Xiaouao Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In CVPR, pages 9243–9252, 2020.

[20] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-based sampling for natural image matting. In CVPR, 2019.

[21] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. ACM TOG, 40(4):1–14, 2021.

[22] Andrey Vovny and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In ICMC, 2020.

[23] Ruei Wang and Michael F Cohen. Optimized color sampling for robust image matting. In CVPR, pages i–8, 2007.

[24] Yang Yang Xu, Yong Du, Wenpeng Xiao, Xuemiao Xu, and Shengfeng He. Faithful extreme rescaling via generative prior reciprocated invertible representations. In CVPR, pages 5708–5717, 2021.
Yangyang Xu receives his Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His current research interests include computer vision, image processing, computer graphics, and deep learning.

Zeyang Zhou is a master’s student in the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing and deep learning.

Shengfeng He (Senior Member, IEEE) is an associate professor in the School of Computer Science and Engineering, South China University of Technology. He obtained B.Sc. and M.Sc. degrees from Macau University of Science and Technology in 2009 and 2011 respectively, and a Ph.D. degree from City University of Hong Kong in 2015. His research interests include computer vision, image processing, computer graphics, and deep learning. He serves on the editorial board of Neurocomputing.