EmoStyle: One-Shot Facial Expression Editing Using Continuous Emotion Parameters

Bita Azari and Angelica Lim
Simon Fraser University
Burnaby, Canada
{bazari, angelica}@sfu.ca

Abstract

Recent studies have achieved impressive results in face generation and editing of facial expressions. However, existing approaches either generate a discrete number of facial expressions or have limited control over the emotion of the output image. To overcome this limitation, we introduced EmoStyle, a method to edit facial expressions based on valence and arousal, two continuous emotional parameters that can specify a broad range of emotions. EmoStyle is designed to separate emotions from other facial characteristics and to edit the face to display a desired emotion. We employ the pre-trained generator from StyleGAN2, taking advantage of its rich latent space. We also proposed an adapted inversion method to be able to apply our system on real images in a one-shot manner. The qualitative and quantitative evaluations show that our approach has the capability to synthesize a wide range of expressions to output high-resolution images.[1]

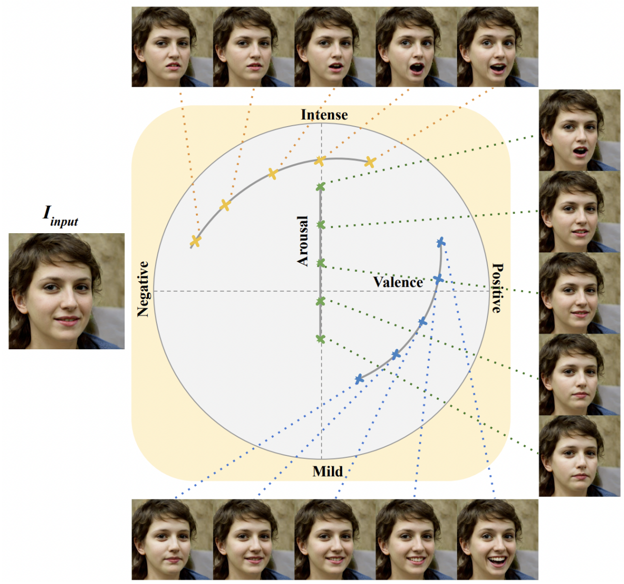


Figure 1. The input image is displayed on the left. The images at the top, right, and bottom of the plot represent the outputs generated by EmoStyle using continuous emotion parameters in the valence and arousal space.

1. Introduction

Facial expression editing is an active research field with applications in areas such as entertainment, virtual assistants, and psychology research. In the field of emotion psychology, scientists need ultra-realism, diversity, and a continuous, scientifically-supported control space, and are eagerly seeking a tool to improve upon WEIRD (Western, Educated, Industrialized, Rich Democracies) real face stimuli, e.g. NimStim [33] and Chicago [20]. Similarly, the visual effects (VFX) community needs a precise emotion editing tool that edits the face only, maintaining all other aspects (e.g. hair, skin tone). The ability to synthesize realistic facial expressions has the potential to enhance human-agent interaction and improve emotional intelligence. Currently, the process of editing facial expressions with high control typically involves creating 3D animated humans, which can be a resource and time-intensive task [5]. Therefore, it is crucial to explore alternative and more efficient methods for synthesizing realistic facial expressions.

The study of emotions has a long history in psychology. In the 1960s, Paul Ekman [10] proposed a widely accepted categorization of facial expressions. He identified six basic emotions: happiness, sadness, anger, fear, surprise, and disgust, which were later expanded to include two additional emotions: contempt and embarrassment. More recently, researchers have focused on the dimensional nature of emotions, with **valence (V)**, which reflects the negativity (-1) or positivity (+1) of emotion, and **arousal (A)**, which reflects the level of physiological activation associated with the experience [30], from mild (-1) to intense (+1). These dimensions have been used to describe a wide range of emo-

[1] https://bihamta.github.io/emostyle/

tions (compared to only 8 categorical facial expressions) and have been incorporated into many models of emotion recognition and synthesis [16, 32]. As an example of the advantages of the dimensional approach, it can distinguish between cold anger and hot anger, and low arousal positive (considered "ideal affect" in East Asian culture) and high arousal positive (ideal affect in North America) [35]. According to a study conducted by Arias *et al*. [2], utilizing Generative Adversarial Networks (GANs) to create slightly more smiling faces can improve both human-human and human-computer interactions.

Early work in 2D facial expression editing employed Conditional Generative Adversarial Networks (GANs) to modify facial images [9, 17, 26]. Such studies also used valence and arousal as editing parameters, yet worked mainly on low-resolution images and tended to produce artifacts on the human faces. In recent years, the StyleGAN/StyleGAN2 [13, 14] models have revolutionized the field of image synthesis and are one of the most widely used generators. In the area of facial expression, studies using StyleGAN2 have primarily focused on creating slight variations in emotional expressions, such as increasing the level of smiling or anger [1, 31]. Therefore, major limitations remain in the ability to edit and synthesize more nuanced and high-quality emotional expressions. Additionally, these StyleGAN2 methods have primarily focused on image editing within the model's existing domain [1, 11, 31], limiting their applicability to unseen faces.

This paper presents an approach for one-shot, high-quality editing of human facial expressions based on valence and arousal, as opposed to categorical facial emotions. Firstly, we contribute a method for disentangling emotion expression from other facial attributes, by training a nonlinear Emotion Extraction module using an alternating emotion variation and emotion reconstruction method. A key insight is that by training our model with a broad range of valence and arousal values, we can increase the diversity in the output facial expressions. Ultimately, we can create facial expressions that are slightly more or less surprised, disgusted, or tired, among others.

Secondly, we propose a combination of auxiliary loss functions aimed at facilitating facial expression editing while preserving other facial attributes. One of the key contributions of our work is a novel background loss function, which ensures that the model preserves skin color, background, and hairstyles in extreme emotional modifications. This is achieved by applying a mask over the face, excluding the forehead, and enforcing the model to maintain consistency in all other areas. By doing so, the model learns to preserve skin color by maintaining consistency in the forehead region, thus enabling more realistic and accurate facial expression synthesis.

Finally, we propose an extension to our facial expression editing method to enable one-shot editing on real images (out-of-StyleGAN2 domain images). To handle faces not seen during initial training, we describe a fine-tuning approach that builds upon previous facial inversion methods [28, 34]. To showcase the effectiveness of our approach on unseen images, we evaluate our method on faces from CelebA [19], widely used as out-of-domain images for StyleGAN2, which was trained on FFHQ [13] comprised of photos from Flickr. The results presented in our study highlight the efficacy and potential of our approach for facial expression editing on real images.
Our approach is capable of producing high-quality images with a resolution of 1024 x 1024 pixels, which is currently the maximum resolution that can be achieved using StyleGAN2. As a result, our proposed method for synthesizing facial expressions based on valence and arousal provides greater flexibility and control over emotional modifications in facial images. This approach allows for more nuanced and subtle modifications to facial expressions, enabling greater realism and accuracy in the synthesized images.

2. Related Work

In recent years, generative models have gained significant attention for their ability to produce realistic images. In addition to their remarkable generation capabilities, generative models can also be utilized for image editing. Here, we review various techniques that aim to alter facial expressions using generative models.

2.1. Facial expression synthesis

One of the first facial editing approaches used conditional GANs [23], which condition image generation on a label. ExprGAN [9] is based on conditional GAN architecture and Adversarial Autoencoders [21] to synthesize emotional expressions. Inspired by ExprGAN, Lindt *et al*. in [18] proposes CAAE [38] for emotion-based expression editing, incorporating identity preservation. As highlighted in their study, they struggle with maintaining identity during extreme emotions. The generated images also have a low resolution of 96x96 pixels. GAN-imation proposed by Pumarola *et al*. [26] employs a version of conditional GANs and utilizes valence and arousal, in addition to categorical emotion labels, to synthesize facial expressions on a face. Another example is StarGAN [6], a conditional GAN that has been modified in VA-StarGAN [17] to allow for face editing based on valence and arousal intensities. However, despite their potential, the generated images often contain artifacts and the expected results may only be achievable on low-resolution images. Differing from traditional approaches reliant on manual labels, d'Apolito *et al*. in GANmut [7] presents a GAN-based framework. It constructs a nuanced interpretable emotional conditional space via fundamental categorical emotion labels. However, the generated images exhibit limitations, particularly in quality notably around the eyebrows and mouth.

2.2. Semantic editing using StyleGAN2

More recently, high-quality generative models such as StyleGAN [13] and StyleGAN2 [14] have been widely used in face editing tasks owing to their expressive and informative latent space. The rich StyleGAN2 latent space provides the ability to separate face attributes from other facial features. Researchers have made significant progress in semantic editing within this domain, which is comprehensively surveyed by Melnik *et al*. [22].

Many studies have investigated moving along a direction in the latent space to identify corresponding changes that result in specific facial attribute modifications (e.g. age, gender, expression). One such example is StyleFlow proposed by Abdal *et al*. [1], which utilizes continuous normalizing flows to learn a semantic mapping between the Z and W spaces. InterFaceGAN proposed by Shen *et al*. [31] utilizes pre-trained classifiers to learn a hyperplane in the latent space, which serves as a separation boundary to identify directions along which specific facial attributes increase or decrease. GANSpace proposed by Härkönen *et al*. [11] is an unsupervised method that employs principal component analysis to identify directions for image editing. Once these directions are identified, GANSpace relies on the user to manually select the most meaningful directions based on the target attribute by observing the generated outputs. All of the editing methods mentioned above have a limitation in that they provide limited control over the output image, allowing users to only increase or decrease an attribute to a certain extent.

To address this limitation in control, StyleCLIP proposed by Patashnik *et al*. [25] enables the manipulation of facial features using only text prompts, utilizing a contrastive language image pre-training (CLIP) [27] model to learn a joint embedding. Latent-2-latent (L2L) proposed by Khodadadeh *et al*. [15] trains a non-linear attribute model capable of controlling the input latent instead of solely moving along the latent space. Despite the increased ability to control facial expression attributes, StyleCLIP and L2L are prone to modifying unwanted attributes such as identity.

2.3. Editing real images in StyleGAN2

Despite the expressiveness of StyleGAN2, editing real images within the StyleGAN2 latent space can be challenging. As a solution, various inversion methods have been proposed, which have recently been surveyed in Xia *et al*. work [36]. The Pivotal Tuning Inversion (PTI) method proposed by Roich *et al*. [29] utilizes an initial latent code as a pivot and then fine-tunes the generator to reconstruct the image while preserving the remaining parts of the latent code. MyStyle proposed by Nitzan *et al*. [24] builds upon PTI and proposes the use of a convex hull to identify a cluster for an identity in the latent code. By mapping an identity to a convex hull, it enables image editing while preserving the identity, which can be used for super-resolution and other editing tasks. However, in order to identify the convex hull, it is necessary to use approximately 100 images captured under diverse conditions to apply changes without compromising identity [24]. Collecting such a diverse dataset can be impractical when attempting to edit the facial expression of a single individual.

3. Method

Our method takes an image as input I_{input} and two emotion parameters (valence and arousal). As shown in Fig. 2, our pipeline consists of an EmoExtract module M, and a pretrained generator model, StyleGan2, G. In phase 1, we train EmoExtract and the upsampling module to learn how to disentangle emotions from other facial attributes. In phase 2, we freeze the EmoExtract and upsampling modules and fine-tune G on the target real face.

3.1. Phase 1: Training EmoExtract

In the initial phase of our methodology, our goal is to train the EmoExtract module to produce the necessary modifications to the input image to generate a face to attain the target emotion. The process is depicted in Fig. 2.

We provide our 3-layered MLP module, EmoExtract, with a latent code representing a face from the W space of StyleGAN2, concatenated with an emotion latent code representing valence and arousal. In order to train our model, we use generated face images alongside their corresponding latent codes from the W space of StyleGAN2. We feed the target valence and arousal (v, a) to an upsample MLP model to map these two numbers to a higher dimension. Then, we concatenate the emotion embedding and the latent code w and feed it into our EmoExtract M, $M(\text{Upsample}(v, a), w) = d$. EmoExtract modifies the original latent code such that the final image is in accordance with the input emotion parameters $G(d + w) \rightarrow I_{\text{output}}$. In each epoch of our training process, we train the EmoExtract network in one of two different ways:

- **Emotion Variation**: We generate two random VA values and use these emotional variations as inputs. Thus, we aim to generate faces depicting emotions that do not frequently appear in our generated dataset. In this part, we use background loss alongside our three main loss functions (emotion loss, identity loss, and pose loss) to assure the preservation of background, hairstyle and skin color, described in Sec. 3.3.
- **Emotion Reconstruction:** Every fifth batch (chosen through trial and error) we feed the input face to a

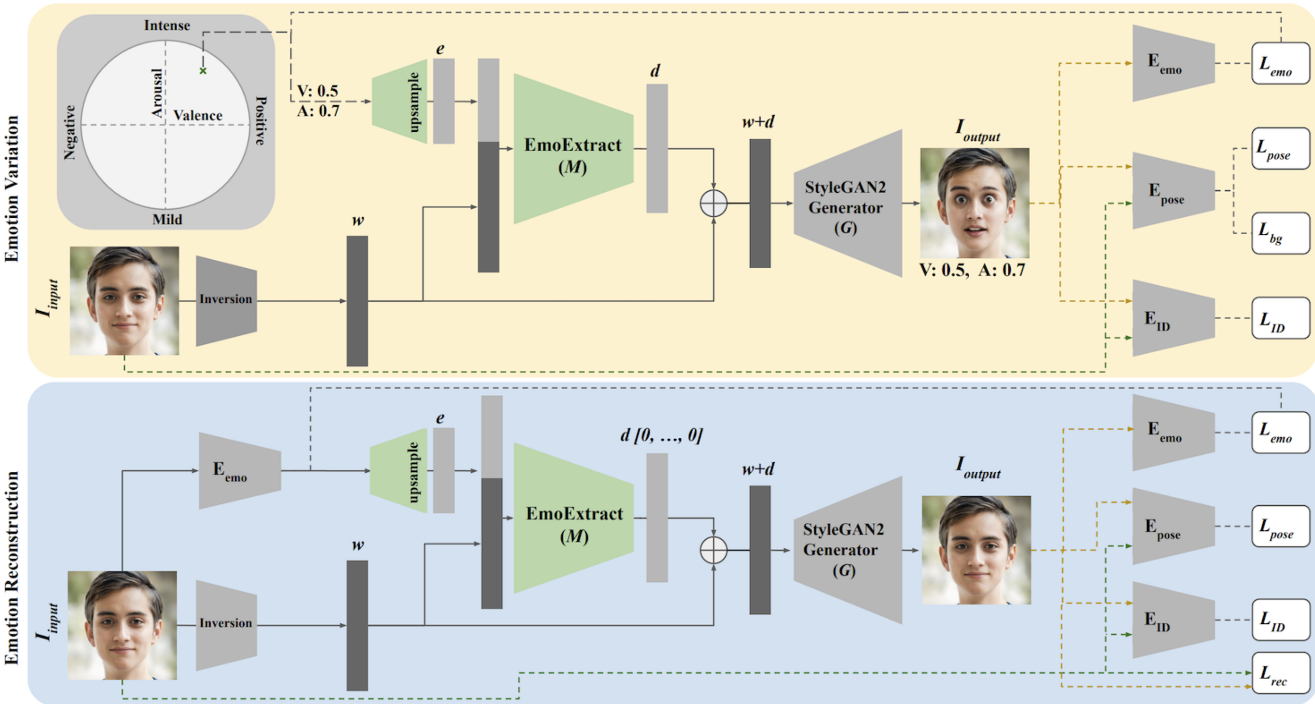


Figure 2. **Phase 1: Training EmoExtract**. We train the EmoExtract and up-sampling modules (green) by alternating Emotion Variation with random emotion parameters from the valence-arousal space (top), with Emotion Reconstruction of the input face (bottom). Five auxiliary losses are used for this purpose, as indicated by the dashed lines. The Inversion module [28] is employed to extract the latent code w of the input image I_{input}. The EmoExtract module is trained to determine the necessary modifications d that should be applied to a latent code w. Note that d should result in 0 for the Emotion Reconstruction segment. The final latent code is generated by adding d to the original latent code w. Finally, the StyleGAN2 generator is used to create our desired image.

state-of-the-art valence and arousal estimation network proposed by Toisoul *et al*. [32] and predict the emotion of the face. Then, we use these emotion parameters of the input face to enhance reconstruction performance and assist the network in producing realistic outputs. We train EmoExtract to produce a zero vector that indicates no adjustment is required between the facial expression and the target emotion. During this process, we incorporate reconstruction loss along with other loss functions (emotion loss, identity loss, and pose loss) to ensure that the output image accurately represents the input image.

By applying such steps during the training process we enforce the EmoExtract model to learn the correlation between the target emotion and the emotion coded within the latent code.

3.2. Phase 2: Fine-tuning StyleGAN2

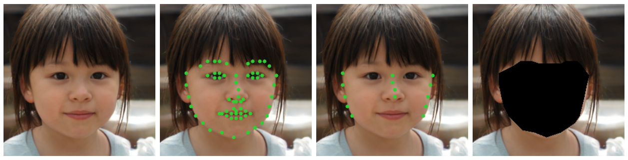
Next, we describe our method for allowing the editing of a new person's face who is out of the StyleGAN2 domain. In this second phase, we freeze the EmoExtract module trained previously and fine-tune our StyleGAN2 component. Our inputs during this phase are emotion parameters and one real face. First, we determine the face's latent code utilizing an inversion framework to extract the latent code in the StyleGAN2 W space, then perform a fine-tuning step inspired from [29].

We incorporate the same loss functions from Phase 1 into the optimization process and fine-tune the generator on a single image. This fine-tuning allows us to accurately reconstruct the input image and grants us the ability to perform edits. In this phase, we fine-tune the StyleGAN2 generator to adjust it in a way that it can move our latent code to a more editable space in the latent space of StyleGAN2.

3.3. Loss Functions

To ensure high-quality reconstruction, accurate facial emotion synthesis, and preservation of identity, pose, and background, we employ a weighted combination of five loss functions:

Emotion Loss ($\mathcal{L}_{\textbf{emo}}$) is employed to assess whether the generated output images reflect the input emotion parameters. This is accomplished by computing an $L2$ loss between the valence-arousal values of the input image and those of the predicted valence-arousal derived from the generated image. We use a pre-trained emotion estimation



(a) (b) (c) (d)

Figure 3. a) Input image, b) Landmarks detected by the model proposed by Bulat and Tzimiropoulos [3], c) Subset of landmarks chosen for calculating pose loss, d) Estimated mask over face to use for background loss.

model to predict valence and arousal [32].

$$\mathcal{L}_{\text{emo}} = \left\| \text{emo}\left(I_{\text{input}}\right) - \text{emo}\left(I_{\text{gen}}\right) \right\|_2 \tag{1}$$

Identity Loss ($\mathcal{L}_{\text{id}}$) is employed to preserve the identity of the input image, we use a state-of-the-art face recognition system (VGGFace2 [4]) and calculate the $L1$ loss between the embeddings of the input and generated images.

$$\mathcal{L}_{\text{id}} = \left\| \text{Em}_{\text{id}}\left(I_{\text{input}}\right) - \text{Em}_{\text{id}}\left(I_{\text{gen}}\right) \right\|_1 \tag{2}$$

Pose Loss ($\mathcal{L}_{\text{pose}}$) is utilized to ensure that the generated image preserves the pose and facial alignment of the input. To achieve this, we apply an $L2$ loss on a subset of face landmarks estimated by a pre-trained Facial Alignment Network (FAN) [3]. The selection of this subset is based on the consideration of landmarks that remain relatively stable despite changes in facial emotions. The 14 selected landmarks are shown in Fig. 3c.

$$\mathcal{L}_{\text{pose}} = \left\| \text{Pose}\left(I_{\text{input}}\right) - \text{Pose}\left(I_{\text{gen}}\right) \right\|_2 \tag{3}$$

Reconstruction Loss ($\mathcal{L}_{\text{rec}}$) is utilized to enforce high-quality image reconstruction. To achieve this, we adopt the "mix" loss approach proposed by Zhao *et al*. [39] which involves a weighted combination of $L1$ loss and MS-SSIM loss.

$$\begin{aligned}\mathcal{L}_{\text{rec}} =& \alpha\left(1 - \text{MS-SSIM}\left(I_{\text{input}}, I_{\text{gen}}\right)\right) \\ &+ (1-\alpha)\left\| I_{\text{input}} - I_{\text{gen}} \right\|_1\end{aligned} \tag{4}$$

Notably, we employ the reconstruction loss on the whole image on Emotion Reconstruction batches where we use the original emotion parameters of the input image as the target, indicating that the generated face is expected to resemble the input image.

Background Loss ($\mathcal{L}_{\text{bg}}$): In the Emotion Variation batches where the valence and arousal are random numbers, we alter the reconstruction loss in Eq. 4 to enforce only the preservation of the hair and background. To accomplish this, a mask is estimated on the facial region using the facial landmarks that are extracted by the Facial Alignment Network (Fig. 3d). The remaining regions of the input and output face are compared using the reconstruction loss, according to Eq. 4, except that I_{input} and I_{gen} are masked.

	LPIPS ↓	FID ↓	ID ↑	VA std ↑
GANSpace [11]	0.47	30.31	0.53	0.5/0.2
InterFace [31]	0.36	11.83	0.87	0.4/0.1
StyleFlow [1]	0.36	13.03	0.83	0.5/0.1
GANmut [7]	0.26	8.25	0.81	0.5/0.25
L2L [15]	0.19	16.19	0.78	**0.5/0.3**
EmoStyle (Ours)	**0.07**	**7.86**	**0.88**	0.5/0.25

Table 1. EmoStyle outperforms other methods in FID, LPIPS, and ID preservation while maintaining high valence and arousal standard deviation. VA std is indicating valence and arousal standard deviation.

The overall auxiliary loss is calculated as a weighted sum of the individual losses. The specific weights are determined through trial and error, by comparing our metric performance using different weight combinations.

$$\begin{cases}\mathcal{L}_{\text{EmoVar}} = \lambda_1\mathcal{L}_{\text{emo}} + \lambda_2\mathcal{L}_{\text{pose}} + \lambda_3\mathcal{L}_{\text{bg}} + \lambda_4\mathcal{L}_{\text{id}} \\ \mathcal{L}_{\text{EmoRec}} = \lambda_1\mathcal{L}_{\text{emo}} + \lambda_2\mathcal{L}_{\text{pose}} + \lambda_3\mathcal{L}_{\text{id}} + \lambda_5\mathcal{L}_{\text{rec}}\end{cases} \tag{5}$$

$\mathcal{L}_{\text{EmoVar}}$ represents the total loss when random valence and arousal are used (Emotion Variation), while $\mathcal{L}_{\text{EmoRec}}$ represents the total loss when emotion modification is not desired (Emotion Reconstruction).

4. Experiments

In this section, we undertake a thorough evaluation of our system. Our evaluation is conducted both quantitatively and qualitatively to ensure a comprehensive analysis of the system. Additionally, we conduct an ablation study to evaluate the effectiveness of each component in our pipeline.

4.1. Experimental Settings

We utilized 70,000 generated images along with their latent codes for training, and a separate set of 1,000 images with their corresponding latent codes for testing. These images were generated using StyleGAN2, which was originally trained on the FFHQ dataset [13]. Following the recommendation in the original StyleGAN paper, we truncated the vectors by a factor of 0.7. For qualitative real-image experiments, we selected a subset of the CelebA dataset [19] that contains high-resolution images of human faces. The proposed pipeline was trained on a system equipped with a GeForce RTX 2080 GPU, using the PyTorch deep learning library. We utilized StyleGAN2 pre-trained at 1024 x 1024 resolution for all our experiments. To compute losses and train our network, in every 5 batches, we employed Emotion Reconstruction which uses the estimated valence and arousal values extracted by our emotion estimation model, and Emotion Variation in the remaining batches. The loss

Figure 4. Sample outputs of EmoStyle, with each output corresponding to a different combination of input valence and arousal. The results align with our expectations, as they effectively convey the positivity or negativity and the level of arousal of the faces.

weights were set empirically to $\lambda_1 = 0.3$, $\lambda_2 = 0.001$, $\lambda_3 = 0.2$, $\lambda_4 = 1.5$, and $\lambda_5 = 0.2$.

4.2. Quantitative Evaluations

We conducted a thorough evaluation of our framework's performance, comparing it with current semantic editing methods in terms of editing quality and identity preservation capabilities. To accomplish this, we employed two distinct types of metrics: the Fréchet distance (FID) [12], Learned Perceptual Image Patch Similarity (LPIPS) [37] and identity preservation. Facial expression edits were performed on 1000 images with different valence and arousal values, and the results were compared with GANmut [7] and those of existing StyleGAN2 methods, including L2L [15], InterFaceGAN [31], StyleFlow [1], GANSpace [11]. VA values chosen for this experiment can be found in Section 7.4 of the Supplementary Material.

FID and LPIPS Scores: These were utilized to measure the diversity and quality of the generated images. Specifically, we reported FID and LPIPS scores for 1000 images generated using StyleGAN2 and our edited images. The results are presented in Table 1.

Identity Preservation Score: To evaluate our framework's identity preservation capability, we employed an external face recognition model ArcFace [8] and calculated the cosine similarity between the original and edited images. We then compared our results with those of existing methods, and the findings are presented in Table 1.

Valence and Arousal Standard Deviation: We compared the diversity of facial expressions in our method with prior work by calculating the standard deviation of valence and arousal. We used a pretrained emotion estimation module [32] to estimate these values for the generated images and presented the results in Table 1. Table 1 demonstrates that

Figure 5. The result of our real image EmoStyle fine-tuning, compared to standard image inversion without fine-tuning using two different methods, e4e [34] and pSp [28]. As shown, both alternative methods fail to preserve the identity of the person.

EmoStyle outperforms prior work in terms of FID, LPIPS, and ID preservation. While L2L demonstrates a slightly higher VA std, it is crucial to understand that a wider range in these values alone may not necessarily signify a more diverse range of emotions; artifacts may also contribute to unintentional modifications of facial expressions (Fig. 6b). We additionally computed the root mean squared error (RMSE) between the target VA values and the corresponding predictions derived from the generated images. Notably, this metric was calculated exclusively for our model and L2L, as other models did not explicitly incorporate VA value calculations. The RMSE for L2L stands at 0.187, while for EmoStyle, it is 0.181.

4.3. Qualitative Evaluations

In terms of qualitative results, we illustrate our sample outputs in Fig. 4, which displays face images generated using different valence and arousal values. We also evaluated the effectiveness of our emotion editing method by comparing it to three existing face editing methods (GANimation, StyleCLIP and GANmut) in Fig. 6a. We focused on four basic emotion categories for visualizing the results. Initially, we tested various text prompts as inputs to Style-

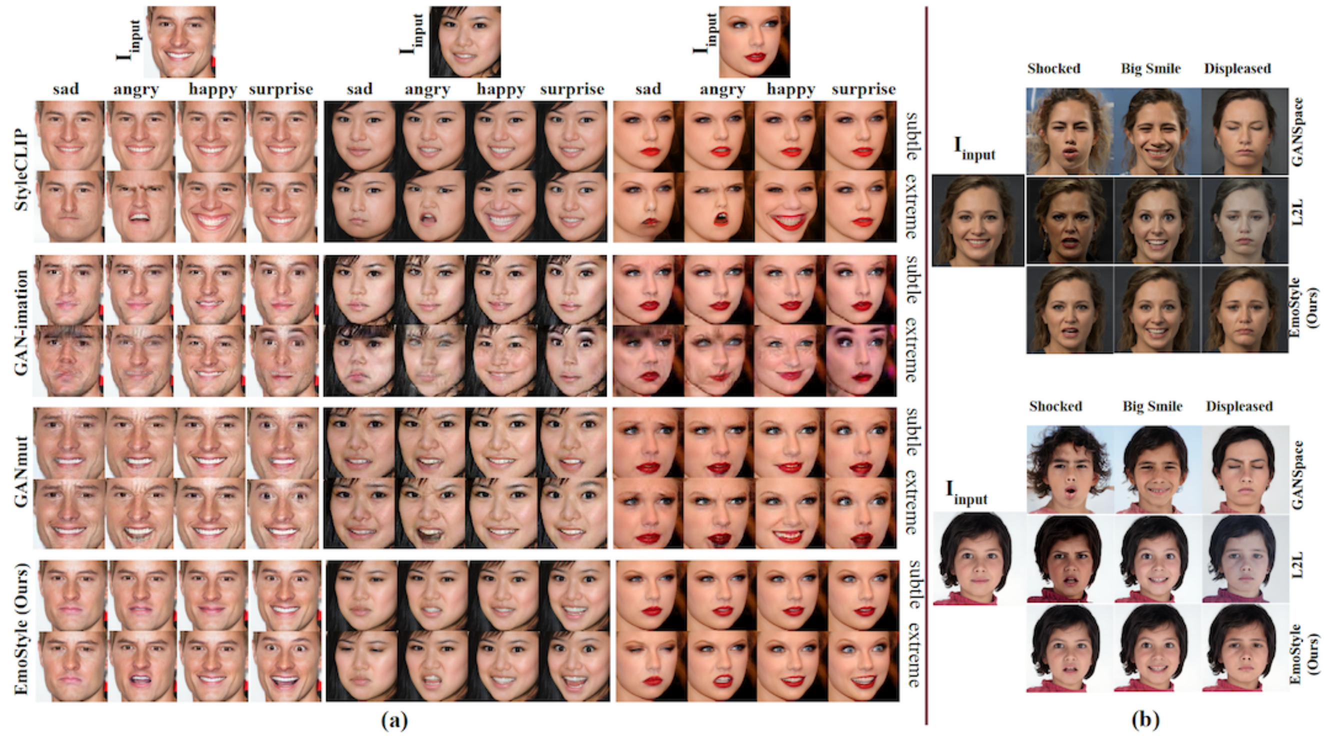


Figure 6. a) StyleCLIP [25], GAN-imation [26], and GANmut [7] are 3 approaches that are closely related to EmoStyle, as they can handle real images with varying degrees of emotion intensity. The images at the top represent the input images, while the subsequent rows of images illustrate different intensities of emotions generated using various methods. As shown in this image, all three struggled to effectively synthesize the desired emotions in a given image and were unable to maintain the identity of the original image. Note that GAN-imation and GANmut generated images at 128 x 128 resolution and others at 1024 x 1024. b) Outputs from GANSpace, L2L, and our system. Both L2L and GANspace models fail to preserve identity in some emotional expressions.

CLIP to generate facial expressions but found that StyleCLIP could only produce a limited set of discrete expressions. To demonstrate the diversity of facial expressions generated by GAN-imation, we used valence and arousal with varying intensities to map to the basic emotions of our choice. We repeated the same procedure with our system to produce the same basic emotions with different intensities. For GANmut, we located the emotion categories within GANmut's personalized latent space. Since GAN-imation can perform on cropped, low-resolution (128 x 128) images, we cropped the face bounding box in order to compare our results with those obtained from StyleCLIP, GANmut, and EmoStyle. Results are shown in Fig. 6a. In order to evaluate the plausibility of our results and compare them to StyleGAN2 methods, we assessed our method against GANSpace and L2L. We selected these two approaches because they enable us to control the emotion of the generated image. StyleFlow and InterFaceGAN do not offer direct control over the output expression, instead relying on relative adding or subtracting in a specific direction, therefore we omitted them for comparison. GANSpace computed control vectors for diverse facial attributes, 10 of which are discrete emotion-related states, such as a big smile and fearful eyes. To compare our method with GANSpace, we used the GANSpace model to synthesize emotional faces based on their annotated attributes. We implemented L2L based on their published paper [15] and used the same emotion estimation instead of their attribute module. We chose the closest valence and arousal values to the labels predefined in GANSpace. We then used EmoStyle and L2L models to synthesize these emotions on the same faces. We notice that both GANSpace and L2L lose identity preservation when changing emotions. The results are presented in Fig. 6b.

4.4. Ablation Study

To demonstrate the effectiveness of each component in our system, we conducted an ablation study by comparing the results obtained from 5 different settings. To underscore the importance of each component, we perform a qualitative and quantitative analysis. We visually compare their impact in Figure 7, and assess their performance through FID and LPIPS metrics, presented in Table 2.

Background Loss: The introduction of a background loss enabled us to exercise control over the image's background in the batches where target emotions were randomly selected. Without the masked background loss, we observed significant variations in the background, skin colour, and hairstyle of the generated faces, as illustrated in Fig. 7.

Identity Loss: To maintain the identity of the face while changing emotions, we integrated an identity loss component into our pipeline. As demonstrated in Fig. 7, our system occasionally failed to preserve the identity of the face when the identity loss was not incorporated into the system.

Emotion Reconstruction: As described in Sec. 3, every fifth batch, we provide the estimated valence and arousal of the input image as input to our model, with the expectation that it would learn to reconstruct the image without altering the emotions. In this experiment, we omitted this step and instead input random valence and arousal at every step. However, as shown in Fig. 7, our experiments demonstrated a noticeable decline in reconstruction quality and disentanglement when the Emotions Reconstruction was not considered in the training process.

Emotion Variation:: In Fig. 7, we also provide visual evidence of the significance of utilizing the Emotion Variation method in our pipeline. When we excluded this step, our model did not learn how to edit the person's emotional expression.

Personalization : We show the effect of fine-tuning StyleGAN2 using our losses. First, we retrieve the corresponding latent code in the StyleGAN2 latent space using one of two different inversion methods, namely, e4e [34] and pSp [28]. While our method described in Sec. 3 employs pSp as the inversion method and maps the latent code to W, when testing e4e we optimize the weights of a stack of MLP networks to edit the face in $W+$. Fig. 5 shows the importance of fine-tuning StyleGAN2 as it enables us to invert faces to the StyleGAN2 domain and edit it while preserving the identity.

	LPIPS ↓	FID ↓
EmoStyle	7.86	0.07
w/o Identity Loss	9.10	0.09
w/o Pose Loss	10.46	0.1
w/o Background Loss	10.9	0.14
w/o Emotion Reconstruction	12.3	0.12

Table 2. Ablation Study comparing component effectiveness through FID and LPIPS metrics.

5. Discussion

During the training and evaluation of our system, we encountered noteworthy observations. Our architecture incorporates state-of-the-art emotion estimation techniques to estimate valence and arousal. In our initial development, we discovered that the valence and arousal estimation module proposed by Toisoul [32] had high performance, but it tended to also focus on non-facial attributes, such as background, hair type, and age, when estimating emotion. For example, when using it to train EmoStyle and attempting to generate a crying face (low valence, low arousal), the resulting face would tend to resemble an infant, or when in a state of bliss (high valence and low arousal), the background would change to a green representing nature. It was to address these issues that we implemented background loss and ID loss. Another notable finding of our study is that the StyleGAN2-generated images lack emotional diversity. The standard deviation of valence and arousal for 70,000 images generated by StyleGAN2 was 0.42 and 0.15, respectively. In contrast, our proposed method demonstrated the ability to generate images with greater diversity, achieving standard deviations of 0.5 and 0.25 for valence and arousal, respectively. This pattern is also observable in a heatmap illustrating the diversity of valence and arousal across generated images (see Sec. 7.1 of Supplementary Material.) In certain images, distinguishing between emotional expressions where the valence and arousal closely align can be challenging. While existing benchmarks primarily center on the VA dimensions, a third dimension, dominance, remains unexplored. A promising direction for future research lies in the incorporation of this additional axis. In future work, we will explore (Action Units) AUs as an extra control axis: Semantic emotion editing could enable global control, while AUs offer local control.

Figure 7. Ablation study: The absence of Emotion Variation results in failure to edit the facial expression, while lack of other modules results in the loss of identity or facial features (e.g. beard).

6. Conclusion

This paper presents a semantic editing system that allows for precise control over the output face's emotional expression. We train an emotion extraction module to identify the latent code that corresponds to the desired emotion parameters and generate a new face image that exhibits the targeted emotion with minimal changes. Our architecture is capable of performing one-shot emotion editing of a given face, even if the face is not present in the latent space of the StyleGAN2 generator. To enable this functionality, we fine-tune the generator with our emotion extraction module. We also demonstrate the effectiveness of our system through various qualitative and quantitative evaluations. Our experimental results demonstrate that our approach is capable of manipulating facial expressions, and preserving identity while generating high-quality images.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021.

[2] Pablo Arias, Catherine Soladie, Oussema Bouafif, Axel Roebel, Renaud Seguier, and Jean-Julien Aucouturier. Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*, 11(3):507–518, 2018.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*, 2017.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[5] Byungkuk Choi, Haekwang Eom, Benjamin Mouscadet, Stephen Cullingford, Kurt Ma, Stefanie Gassel, Suzi Kim, Andrew Moffat, Millicent Maier, Marco Revelant, et al. Animatomy: an animator-centric, anatomically inspired system for 3D facial modeling, animation and transfer. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[7] Stefano d'Apolito, Danda Pani Paudel, Zhiwu Huang, Andrés Romero, and Luc Van Gool. GANmut: Learning interpretable conditional space for gamut of emotions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 568–577, 2021.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[9] Hui Ding, Kumar Sricharan, and Rama Chellappa. ExprGAN: Facial expression editing with controllable expression intensity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

[10] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[11] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable GAN controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

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

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

[15] Siavash Khodadadeh, Shabnam Ghadar, Saeid Motiian, Wei-An Lin, Ladislau Bölöni, and Ratheesh Kalarot. Latent to Latent: A learned mapper for identity preserving editing of multiple face attributes in StyleGAN-generated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3184–3192, 2022.

[16] Dimitrios Kollias. ABAW: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022.

[17] Dimitrios Kollias and Stefanos Zafeiriou. VA-StarGAN: Continuous affect generation. In *Advanced Concepts for Intelligent Vision Systems: 20th International Conference, ACIVS 2020, Auckland, New Zealand, February 10–14, 2020, Proceedings 20*, pages 227–238. Springer, 2020.

[18] Alexandra Lindt, Pablo Barros, Henrique Siqueira, and Stefan Wermter. Facial expression editing with continuous emotion labels. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[20] Debbie S Ma, Justin Kantner, and Bernd Wittenbrink. Chicago face database: Multiracial expansion. *Behavior Research Methods*, 53:1289–1300, 2021.

[21] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[22] Andrew Melnik, Maksim Miasayedzenkau, Dzianis Makarovets, Dzianis Pirshtuk, Eren Akbulut, Dennis Holzmann, Tarek Renusch, Gustav Reichert, and Helge Ritter. Face generation and editing with StyleGAN: A survey. *arXiv preprint arXiv:2212.09102*, 2022.

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

[24] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. MyStyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022.

[25] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

[26] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. GANimation:

Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[28] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in Style: a StyleGAN encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[29] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022.

[30] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[31] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[32] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021.

[33] Nim Tottenham, James W Tanaka, Andrew C Leon, Thomas McCarry, Marcella Nurse, Todd A Hare, David J Marcus, Alissa Westerlund, BJ J Casey, and Charles Nelson. The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry research*, 168(3):242–249, 2009.

[34] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *arXiv preprint arXiv:2102.02766*, 2021.

[35] Jeanne L Tsai. Ideal affect: Cultural causes and behavioral consequences. *Perspectives on Psychological Science*, 2(3):242–259, 2007.

[36] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. GAN inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

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

[38] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.

[39] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.