

# Emotion amplification of facial videos using a fine-tuned StyleGAN

Yukun Xu<sup>1</sup>, Justin N. M. Pinkney<sup>2</sup>, Yong-Liang Yang<sup>3</sup>, Tianjia Shao<sup>1</sup> (✉), and Kun Zhou<sup>1</sup>

© The Author(s) 2025.

**Abstract** The ability to exhibit appropriate emotions is crucial for the expressiveness and attractiveness of facial videos. However, it is difficult to control the level of emotion, even for experienced actors and amateur podcasters on social networks. In this study, we aim to solve the novel problem of semantically amplifying the emotions of a facial video. This poses new challenges for effectively editing a sequence of video frames in terms of face semantics, emotion adaptiveness, and temporal coherence. Our approach is based on semantic face editing in the disentangled latent space of a state-of-the-art StyleGAN model. We presented a new face dataset with diverse emotions to fine-tune the pre-trained StyleGAN and improve the expressiveness of its original emotion-biased latent space. An emotion-editing subspace was constructed to allow adaptive emotion amplification while preserving other facial attributes. We further propose an effective stitching-tuning technique to ensure temporally coherent video frames. Our work results in plausible emotion amplification for a wide range of facial videos. Qualitative and quantitative evaluations demonstrated the advantages of our method over other baseline methods. The proposed dataset and research code will be made publicly available.

**Keywords** emotion amplification; facial video; semantic editing

## 1 Introduction

Facial videos are becoming increasingly popular due

to rapid advances in social media, especially short-form video-sharing services (e.g., TikTok, YouTube Shorts, Instagram Reels), as well as the continuous development of the TV and film industry. The level of emotion exhibited is a key factor for facial videos to be attractive to viewers. However, in practice, this is difficult to control. Even experienced actors, let alone amateur podcasters, require several rounds of action to deliver satisfactory emotion. A common issue is that video sequences are not sufficiently emotional because of their relatively close-to-neutral facial expressions. Therefore, an easy-to-use emotion amplification method is required to benefit not only TV and film post-production, but also social media content creators.

However, emotion amplification of facial videos poses several technical challenges. First, emotions are demonstrated through facial expressions, which are sophisticated facial movements involving almost all the components of the human face, including the mouth, eyes, cheeks, and forehead. This requires facial editing, influencing various facial components in a *semantic* manner. Second, the amplification process must be *adaptive* to individual frames that are likely to vary among different emotions. Effective representation of emotions and their magnitudes is a worthwhile problem to be solved. Third, the entire facial video must be edited in a *coherent* manner to avoid visible artifacts across frames.

In this study, we devise the first emotion amplification method for facial videos to address these challenges (see Fig. 1). Our method is based on the recent surge in semantic face editing based on StyleGAN and its follow-ups [1–3]. Although existing methods such as Refs. [4, 5] allow the semantic editing of faces with an exceptional level of fidelity to the original, most of them focus on single facial images. Ref. [6] attempts to extend prior

1 State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou 310058, China. E-mail: Y. Xu, xuyk98@zju.edu.cn; T. Shao, tjshao@zju.edu.cn (✉); K. Zhou, kunzhou@acm.org.

2 Lambda, Inc., San Francisco, CA 94107, USA. E-mail: justinpinkney@gmail.com.

3 Department of Computer Science, University of Bath, Bath BA2 7AY, UK. E-mail: y.yang@cs.bath.ac.uk.

Manuscript received: 2023-08-01; accepted: 2023-11-11



**Fig. 1** Our emotion amplification method enables semantic editing of input facial videos, adapting to emotions in individual frames. This results in more expressive faces with temporal coherence throughout the entire sequence, as shown in the bottom row.

work from single images to video sequences and achieves plausible results. However, it only allows editing along a specific direction (for example, smile) without adapting to emotional variations throughout a video sequence. We chose a StyleGAN-based approach, following the method outlined in Ref. [6], to perform semantic face editing on a video sequence while addressing the remaining two challenges. To adaptively amplify the emotions in individual frames, we created an emotion-editing subspace within the StyleGAN latent space, enabling a comprehensive and quantitative representation of different emotions and their intensities. To overcome the biased emotion distribution of the original StyleGAN trained on the FFHQ dataset [1] (mainly neutral and happy faces from Flickr users), we further propose a new face dataset comprising 186,553 images collected from movie trailers with diverse emotions shown by actors. Fine-tuning StyleGAN with the new dataset significantly enhances the expressiveness of the emotion-editing subspace, thus leading to better emotion amplification results. In addition, we improved the temporal coherence of the resultant video using a carefully designed stitching procedure, avoiding unwanted visual artifacts due to shape and lightness variations between neighboring frames.

We evaluate our method extensively using a variety of video sequences collected from films, TV shows, and cartoons. These sequences contained faces of varying genders, ages, ethnicity, and emotions. We also propose specific quantitative metrics to measure the quality of emotion amplification. Both the quantitative and qualitative results demonstrate the advantages of our method over baselines for generating satisfactory emotion amplification results.

The main contributions of our work are as follows: First, we present the first emotion amplification method that can generate consistent, adaptive, and

coherent results throughout an entire video sequence. Second, we constructed a new Trailer Faces High Quality (TFHQ) dataset representing a large and emotionally diverse collection of high-resolution face images, allowing unbiased emotion editing in the fine-tuned StyleGAN latent space. Third, we propose a temporally coherent face-stitching technique that yields satisfactory results without visible artifacts.

## 2 Related work

**StyleGAN-based face editing.** StyleGAN [1] and its variants [2, 3] have garnered significant research interest for their capacity to produce high-quality images within disentangled and semantically meaningful latent spaces. The versatile nature of StyleGAN’s intermediate latent space has led to a wealth of research on effective means of editing and manipulating this space to enable effective image editing and manipulation using only a pre-trained StyleGAN. Typical approaches sought to determine a semantically meaningful latent-space direction along which to perform face editing. Supervised methods use a pretrained classifier for the attribute of interest to determine how to move in the latent space to affect changes in the attribute [7–9]. Other studies explore a variety of unsupervised [10, 11] and self-supervised [12, 13] methods to determine potentially interesting latent directions. Recent advances in highly capable pre-trained multimodal models (namely, CLIP [14]) have led to a multitude of efforts to effectively leverage multimodal models to determine edit directions for almost all arbitrary attributes specified via text [15–20].

Editing real-world images using StyleGAN requires a latent representation that can closely reproduce the original image before the latent manipulation can be applied. Many methods have been developed

to perform the task of encoding an image into the StyleGAN latent space from optimization [2, 21] to training encoder networks [22, 23]. While the latent space of a pre-trained StyleGAN [21] can accommodate highly diverse images, it has been demonstrated that encodings faithful to the original image often lack editability. Recent work has shown that fine-tuning or modifying the weights of a pretrained model is beneficial for maintaining editability and fidelity [4, 5, 24]. Ref. [25] explores the channel-wise space of style parameters in StyleGAN and demonstrated that it is particularly disentangled and suitable for editing compared to other intermediate latent spaces. In this study, we utilized StyleGAN to edit individual facial video frames. However, we refined the original StyleGAN latent space to make it less emotion biased for better emotion amplification.

**Video generation/editing using StyleGAN.** Compared to the image domain, StyleGAN-based video generation and editing has been explored less frequently. MoCoGAN-HD [26] trains an LSTM-based generator that can generate a latent motion trajectory, which is used by the pre-trained StyleGAN generator to synthesize the content of individual video frames. StyleVideoGAN [27] embeds training videos in the StyleGAN latent space and trained a temporal structure on latent code sequences. The training process can be based on only one video, and the generated motion information can be transferred to different subjects. StyleGAN-V [28] concatenates motion codes to the constant input tensor of the StyleGAN generator to generate videos of arbitrary lengths and frame rates. It requires temporal data to learn motion information, and only the first frame can be edited and animated without semantic editing control of the entire sequence. PV3D [29] incorporates 3D representations into StyleGAN structures to generate 3D-aware portrait videos. Oorloff and Yacoob [30] encoded the identity, pose, and expressive facial attributes of a video to facilitate its reenactment. Yao et al. [31] investigated the manipulation of facial attributes in StyleGAN latent space for facial video editing. A carefully designed pipeline was proposed to crop, align, invert, edit, and paste faces. For face editing, an attribute-specific latent transformer is trained based on a pre-trained multi-attribute classifier to reach the target attribute while preserving other attributes. Tzaban

et al. [6] followed a similar workflow, emphasizing the preservation of global coherence and preventing inaccuracies in StyleGAN inversion during the editing process. Editing is performed directly on each video frame based on existing image editing tools [7, 15] without training any modules. Jiang et al. [32] disentangled the StyleGAN latent space, propagating face editing from one frame to an entire video sequence while preserving the original motion, edited identity, and temporal coherence. While Refs. [6, 31, 32] demonstrate plausible results, they uniformly applied editing to all frames in the original StyleGAN latent space. However, this approach falls short of meeting our requirements for emotion amplification, which needs to be expressive for various emotions and adaptive to individual video frames

**Image-based emotion analysis.** Emotion analysis is a well-known multimodal problem in which signals with different modalities (e.g., audio, image, video, shape, and animation) can be used to infer the underlying emotion [33, 34]. Emotion recognition using two-dimensional (2D) images is an active research topic in the field of computer vision. Traditional methods primarily use handcrafted features [35–39] for recognition. With the availability of large-scale labelled datasets [40–42], deep learning-based recognition models are preferred [43–47]. The end-to-end nature of such models can also enhance image-based face reconstruction by incorporating emotional consistency loss [48]. In this study, we are more interested in semantically amplifying emotions than recognizing them. Our method directly manipulates image content without fitting/reconstructing a 3D parametric face model, such as the 3DMM [49] which is complex (we use it only for evaluation). A recent study that shares a similar inspiration is Ref. [50], which performed tensor decomposition in the (emotion-biased) StyleGAN latent space to obtain editing directions. However, emotion editing is restricted to the image domain under one specific editing direction (e.g., happiness or sadness), whereas our goal is to amplify the emotions of all facial video frames.

### 3 Method

Given an input facial video with  $M$  frames ( $\{I_i\}_{i=1}^M$ ), our goal is to enhance the emotion of each frame  $I_i$  without introducing temporal artifacts or undesired

alterations in unrelated attributes such as identity, pose, and background. Ref. [6] proposes a semantic facial video editing pipeline based on StyleGAN, which is able to extract and edit faces along a specific editing direction in the latent space, and stitch the edited faces back into the input video frame. While editing is performed on each frame separately, the resultant video is of high quality, owing to the specifically designed encoder for smooth inversion and the fine-tuned generator that allows semantic editing across the frame sequence. We followed this pipeline and contributed to editing and stitching steps (Fig. 2). Instead of uniformly editing all the frames using a single editing direction  $d$ , we adaptively amplified the emotions of individual frames while preserving other attributes. This relies on an innovative approach that accurately models the emotion-editing space  $\mathcal{E}$  (Section 3.1) and dynamically estimates the editing direction  $d_i$  within this space for emotion amplification in each frame  $I_i$  (Section 3.2). We also devised an enhanced stitching technique that eliminates temporal artifacts and results in coherent video frames (Section 3.3). We elaborate the technical details of our method as follows.

### 3.1 Emotion editing space formation

#### 3.1.1 Emotion editing basis

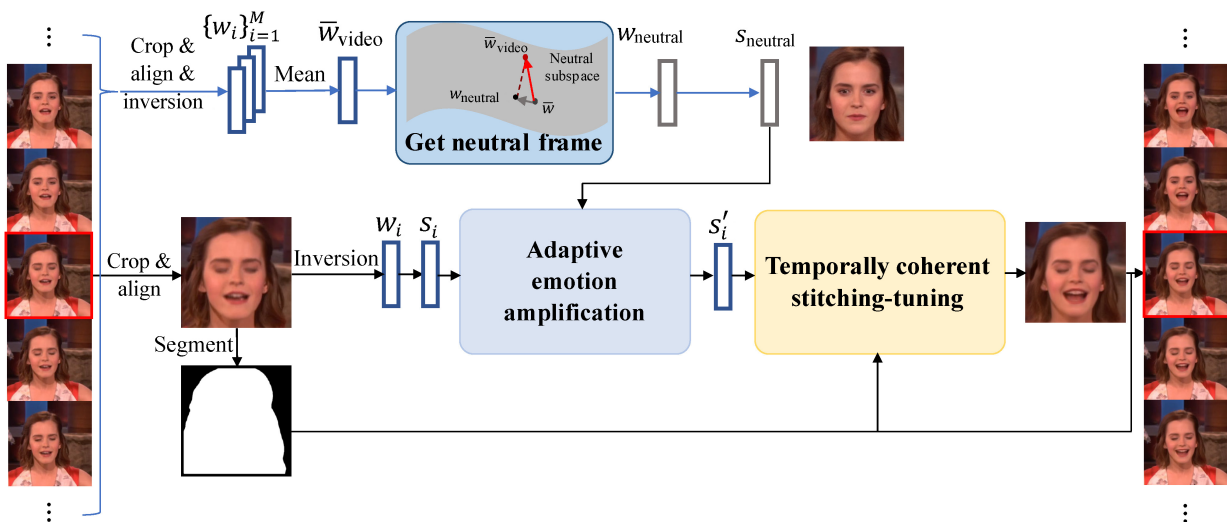
For adaptive and plausible emotion amplification, we extract a set of basis emotion editing directions  $\{e_i\}_{i=1}^7$ , each designed to enhance one of the seven

prototypical emotions: *happiness*, *sadness*, *surprise*, *fear*, *disgust*, *anger*, and *contempt* [42]. We employed StyleCLIP [15] to extract  $\{e_i\}_{i=1}^7$  as global editing directions in StyleGAN's style space  $\mathcal{S}$  [25] for more distinctive and disentangled manipulation. As  $\mathcal{S}$  is affinely transformed from StyleGAN's latent space  $\mathcal{W}+$ , using learned affine transformation for each layer of the generator of StyleGAN, it is trivial to convert  $w \in \mathcal{W}+$  into its counterpart  $s \in \mathcal{S}$ . The emotion-editing space  $\mathcal{E}$  is then defined as a subspace of  $\mathcal{S}$  spanned by all basis editing directions.

However, some of the basis editing directions extracted from the pre-trained StyleGAN were not effective in amplifying certain prototypical emotions, such as disgust and fear (see Fig. 3). We attribute such effects to the biased nature of StyleGAN latent space trained on the FFHQ dataset, which mainly contains neutral and happy faces and lacks samples of faces with other emotions (see Fig. 4(b)). To mitigate this problem, we proposed fine-tuning the existing FFHQ StyleGAN model on a more emotionally varied facial dataset.

#### 3.1.2 Improving the emotions of StyleGAN

To find a dataset of high-quality facial images showing a wider range of emotional expressions, we targeted high-definition movie trailers as the data source. We collected a large dataset of high-resolution face images, called TFHQ, with better coverage of emotions to fine-tune the StyleGAN

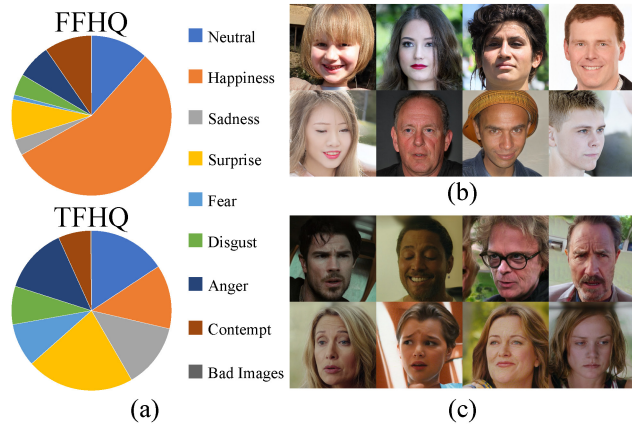


**Fig. 2** The pipeline of our emotion amplification method for editing facial videos. Given an  $M$ -frame input video  $\{I_i\}_{i=1}^M$ , for each frame  $I_i$ , we adopt the approach in Ref. [6] to crop/align facial region  $r_i$  and invert it into the StyleGAN latent space  $\mathcal{W}+$ , resulting in its latent code  $w_i$ . We synthesize a neutral face  $w_{\text{neutral}}$  from the whole sequence  $\{w_i\}_{i=1}^M$ , and conduct emotion amplification for the  $i$ -th frame with the neutral face in the style space  $\mathcal{S}$ . Finally, the edited face is further optimized and interpolated using an enhanced stitching-tuning technique to ensure temporal coherence when stitching back to its original frame.





**Fig. 3** Our method generates more expressive prototypical emotion editing directions. For the same input face, we show the results based on the original StyleGAN-FFHQ (left) and our fine-tuned StyleGAN-TFHQ (right).



**Fig. 4** (a) Our fine-tuned StyleGAN-TFHQ model can generate faces with a much more balanced emotion distribution than the original StyleGAN-FFHQ model. (b) and (c) show randomly sampled faces from the original StyleGAN-FFHQ, and our fine-tuned StyleGAN-TFHQ, respectively.

model. The TFHQ was collected by downloading all movie trailers and features listed on the Apple Movie Trailers website as of August 2022. These 15,379 trailers were downloaded at Full HD (1080p) resolution, amounting to approximately 2 TB/507 h of video. We utilized the pre-trained Yolov5-face large model [51] for face detection in each frame. Faces with a bounding box height less than 256 pixels or a confidence score below 0.5 were excluded. We filtered and selected the sharpest among similar consecutive frames, using image similarity measured with a pre-trained CLIP ViT-B/32 [14] and sharpness measured with the variance of the Laplacian. Finally, the images were aligned using the same procedure as that used for FFHQ [52]. As many of the final images exhibit undesirable properties such as occlusion,

non-photographic faces, or overlaid text, we further perform quality filtering by training a classifier on several hundred subjectively determined “good”/“bad” example images, and use this to exclude predicted “bad” images leaving a total of 186,553 images in the dataset. Examples of the facial images are shown in Fig. 5. The full dataset is available for download at [https://github.com/ArgentRaider/Emotion\\_Amplification\\_StyleGAN\\_TFHQ](https://github.com/ArgentRaider/Emotion_Amplification_StyleGAN_TFHQ). As demonstrated by the distribution of the prototypical emotions in Fig. 4 (detailed in the Electronic Supplementary Material (ESM)), TFHQ is a more emotion-balanced dataset than FFHQ.

Using TFHQ, we fine-tuned the original FFHQ StyleGAN2 model, achieving a Fréchet inception distance (FID) [53] score of 2.68. This model continues to generate high-quality face images but is able to represent a much more diverse set of emotions than the FFHQ model (see Fig. 4(c)). Then, we extracted new basis-emotion editing directions in the improved latent space, as before. Improved basis directions can better amplify prototypical emotions (in particular, biased emotions in the original StyleGAN), as shown in Fig. 3.

### 3.2 Adaptive emotion amplification

#### 3.2.1 Neutral face generation

To adaptively perform emotion amplification in the editing space, we first obtain a face with a neutral expression, called the *neutral face*. The facial emotion in one video frame can then be amplified relative to the neutral face. As the input video clip may lack a frame with a neutral facial expression, we suggest synthesizing a neutral face for it.

The basic idea is to construct a neutral face subspace  $\bar{\mathcal{W}}^+$  in the StyleGAN latent space  $\mathcal{W}^+$  and use  $\bar{\mathcal{W}}^+$  to filter out non-neutral expression



**Fig. 5** Randomly selected samples from the TFHQ dataset.

components. To achieve this, we first collected 1000 images (randomly generated by the fine-tuned StyleGAN-TFHQ model) containing a neutral facial expression, as judged by an off-the-shelf emotion classifier [54]. We then used e4e [55] to encode these images in the  $\mathcal{W}+$  space. To construct the neutral face subspace  $\bar{\mathcal{W}}+$ , we calculated the mean  $\bar{w}$  of these images and extracted the first 50 basis vectors for each layer using principal component analysis (PCA) for each layer. For a given video, we employ e4e to encode all frames and calculate their mean  $\bar{w}_{\text{video}}$ , representing the average facial expressions across all input frames. To filter out the non-neutral expression of  $\bar{w}_{\text{video}}$ , we project it onto the neutral face subspace  $\bar{\mathcal{W}}+$ . This is done by first computing the offset vector  $\Delta w = \bar{w}_{\text{video}} - \bar{w}$ , and then projecting this offset using the PCA basis of  $\bar{\mathcal{W}}+$ , resulting in  $\Delta w'$ . The final neutral face  $w_{\text{neutral}}$  for the input video in  $\bar{\mathcal{W}}+$  is computed as Eq. (1):

$$w_{\text{neutral}} = \bar{w} + \Delta w' \quad (1)$$

Figure 6 presents examples of generating a neutral face from an input facial video.

### 3.2.2 Facial emotion amplification

Now, we can amplify the facial emotions of individual video frames according to the neutral face  $w_{\text{neutral}}$  and its counterpart  $s_{\text{neutral}}$  in style space  $\mathcal{S}$ . For the

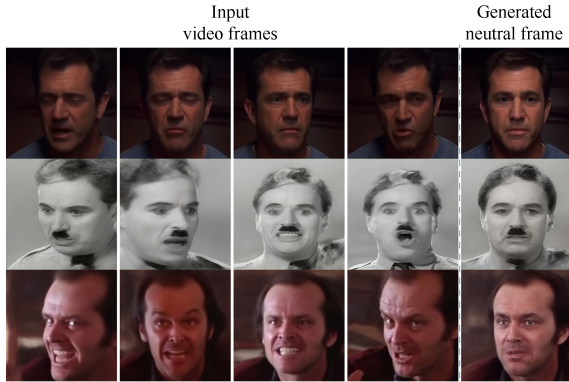


Fig. 6 Example neutral faces generated from input facial videos.

face region  $r_i$  extracted from the  $i$ th video frame, we first use e4e to obtain its position  $s_i$  in the style space. Then, as shown in Fig. 7, we project the offset vector  $\Delta s_i = s_i - s_{\text{neutral}}$  onto the emotion-editing subspace  $\mathcal{E}$ , resulting in an adaptive emotion-amplification direction  $d_i$ . The projection procedure is described as Eqs. (2) and (3):

$$e_j^* = \begin{cases} \frac{e_j}{\|e_j\|}, & j = 1 \\ \frac{e_j - \sum_{k=1}^{j-1} e_j \cdot e_k^*}{\|e_j - \sum_{k=1}^{j-1} e_j \cdot e_k^*\|}, & j > 1 \end{cases} \quad (2)$$

$$d_i = \sum_{j=1}^7 \Delta s_i \cdot e_j^* \quad (3)$$

Finally, the emotion amplification result is obtained as

$$s'_i = s_i + \alpha d_i \quad (4)$$

where  $\alpha$  denotes the amplification magnitude. The final edited image is synthesized in the style space of StyleGAN-TFHQ. Figure 8 displays emotion amplification results for various magnitudes on sampled video frames.

### 3.3 Temporally-coherent stitching tuning

Following Ref. [6] (STIT), the edited face must be injected back into the original video frame as the final step in our pipeline. Instead of performing direct Poisson blending [31], which are prone to inconsistencies around the face border, Ref. [6] proposed a *stitching tuning* technique. This method involves a brief fine-tuning of the face generator to restore the original boundary and preserve the edited face, resulting in visually appealing video frames after stitching. However, as shown in Fig. 9 (middle column), the temporal coherence of the video frames can still be affected, resulting in unwanted artifacts, such as shape/lightness flickering across frames. This is because neighboring frames are achieved separately based on irrelevant generators because fine-tuning is performed on a per-frame basis.

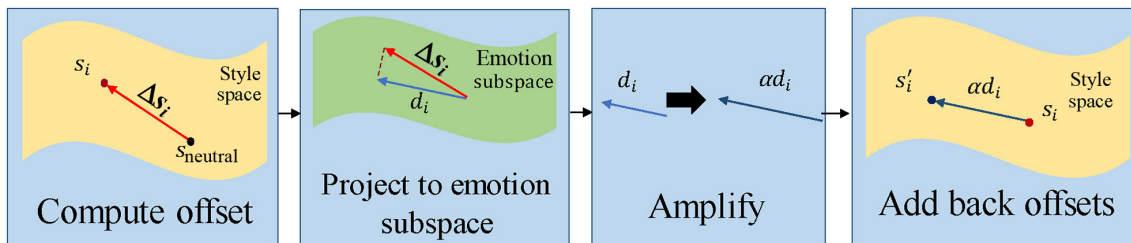
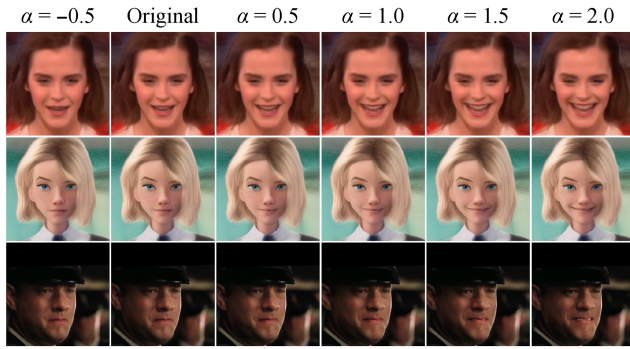


Fig. 7 Amplification stage. The offset latent vector between each frame and the neutral frame is amplified in the emotion subspace.





**Fig. 8** Sampled frames from three facial videos were amplified using our method with different magnitude  $\alpha$ .



**Fig. 9** Comparison between previous stitching tuning (STIT) [6] and ours. Given two neighboring frames in the left column, (a) and (b) demonstrate typical artifacts such as shape flickering (unwanted shape profile change) and lightness flickering (big color difference) artifacts caused by the former, while our work can ensure temporally coherent results in the right column. See the demo video for better visual effects.

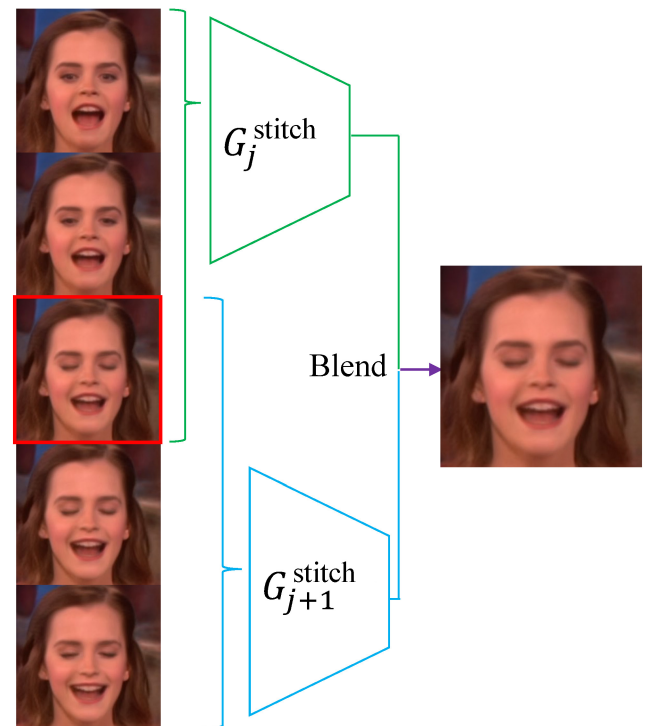
To address this, we propose a *temporally-coherent stitching tuning* technique that fine-tunes the generator across neighboring frames, rather than using only one frame. This helps improve temporal coherence, as the generators are now temporally correlated by training with shared frames along the timeline. In this manner, one frame can be involved in training multiple generators (the number is determined by the neighborhood size). The final

stitching result per frame was then obtained by interpolating the optimized results from all relevant generators of that frame, as shown in Fig. 10. In practice, a small neighborhood size that equals 1 (i.e., only the current frame and its two adjacent frames are used for fine-tuning) can achieve satisfactory results, as shown in Fig. 9 (right).

A more straightforward approach involves tuning the generator for all the frames simultaneously. However, we observed significant losses in image sharpness and detail when applying this tuning approach to a long video sequence (see Fig. 11). This is attributed to the limited expressiveness of the generator of StyleGAN2.

### 3.4 Implementation details

Our experiments were conducted based on StyleGAN2 [2] which we refer to as StyleGAN in this study for simplicity. We fine-tuned the StyleGAN-FFHQ model for a further two million images on TFHQ and used the same hyperparameters as the original model for fine-tuning. Based on StyleGAN-TFHQ, we trained an e4e encoder for the GAN inversion using the same hyperparameters as



**Fig. 10** Stitching tuning stage: To uphold temporal coherence and retain image details, we fine-tune the generator using a limited set of adjacent frames and interpolate the results from the relevant generators for each frame.



**Fig. 11** Comparison between stitch-tuning all frames simultaneously and our approach. (a) and (b) demonstrate significant losses in sharpness and details of the image caused by the former, while our approach can maintain details during tuning.

Ref. [55]. More training and implementation details can be found in the ESM.

We referred to StyleCLIP [15] for retrieving the global emotion directions. When retrieving a global direction, StyleCLIP requires a text prompt to describe the neutral class and another prompt to indicate the attribute of the direction. In our experiments, we use “face” as the text prompt for the neutral class, and use “happy face”, “sad face”, “surprised face”, “terrified face”, “disgusted face”, “angry face”, and “contemptuous

face” for retrieving the global directions of different prototypical emotions.

## 4 Evaluation

In this section, we evaluate our work extensively. We first present the results for a variety of facial video sequences with different genders, emotions, background environments, etc. Then, we present qualitative and quantitative comparisons with baselines to demonstrate the superiority of our method. The test video sequences were either adopted from Ref. [6] or collected from YouTube. To clearly demonstrate the emotion amplification effects, we show only the face and its surrounding area after stitching back the facial region; the corresponding full video frames can be found in the ESM.

### 4.1 Results

Figures 1 and 12 show the final results of our study based on StyleGAN-TFHQ. Additional results can be found in supplemental documents and videos. Our method can handle challenging cases with diverse emotions across different videos and even within a single video. Emotions can be amplified without introducing visible artifacts. Other attributes, such as face pose, identity, skin color, hairstyle, and



**Fig. 12** Emotion amplification results on real facial videos using our method. For each test case, the first row shows sampled frames from the original video, while the second row shows the corresponding frames after emotion amplification. Our method can generate natural and plausible amplification results that can potentially be used for post-production.



background details, can be preserved. Interestingly, by specifying the magnitude of negative emotion amplification  $\alpha$  in Eq. (4) as shown in Fig. 8, our method can achieve emotion suppression (see also the video demonstration). The computational costs of each stage of the proposed method are listed in Table 1.

## 4.2 Comparisons

### 4.2.1 Setup

As there has been no prior work on emotion amplification of facial videos, we compared our results with two baseline methods. The first baseline (Baseline-1) magnifies the face latent code  $w_i$  of the  $i$ th video frame as  $w'_i = w_i + \beta(w_i - w_{\text{neutral}})$ , where  $w_{\text{neutral}}$  is the latent code of the neutral face generated in Section 3.2.1. Note that we magnify only  $w_i$  in the middle layers (5 – 10) of the  $\mathcal{W}+$  latent space, and these layers are the most relevant to the facial expression, whereas the lower (1 – 4) and higher layers (11 – 18) can easily affect the global attributes (e.g., face pose) and local attributes (e.g., face texture), respectively. The second baseline (Baseline-2) follows the same pipeline as our method but in the original StyleGAN-FFHQ latent space. Our method is based on the proposed StyleGAN-TFHQ fine-tuned using the TFHQ dataset. For a fair comparison between different methods, we fix the editing parameter  $\alpha$  at 2.0 for both Baseline-2 and our method, ensuring consistent editing with only the latent space varying. For Baseline-1, we set  $\beta$  to 1.5, aligning the expression magnitudes with Baseline-2 and our method. To mitigate depth ambiguity, we utilize a 3D face reconstruction tool based on the FLAME parametric face model [48] for the resulting video sequence.

### 4.2.2 Qualitative results

Figures 13 and 14 present qualitative comparisons between the proposed and baseline methods. Baseline-1 produces enormous artifacts during amplification, as editing is simply performed on the truncated

latent space, which can easily cause shape distortion and identity changes under the same amplification magnitude. Because of the limited capacity of the original StyleGAN to represent other emotions, such as sadness, Baseline-2 yields unnaturally oversized mouths with non-negligible artifacts on the teeth, as highlighted in Fig. 13. Furthermore, Baseline-2 can introduce other unexpected changes during amplification such as the inconsistent gaze directions shown in Fig. 14 (gazes are clearly shifted in the first, fourth, and fifth columns). To further validate the limited emotion representation capability of the original StyleGAN, we conducted an experiment that amplified the emotions of the entire video in a single



**Fig. 13** Our method generates plausible emotion amplification results without artifacts introduced by baseline methods, such as visible face distortions and identity changes of Baseline-1, and oversized mouth and teeth artifacts (highlighted in the red boxes) of Baseline-2.



**Fig. 14** Our method results in more faithful amplification without unwanted changes of distorted face shape and texture as in Baseline-1, and eye gaze shifts of Baseline-2 (columns 1, 4, 5).

**Table 1** Computational costs of 3 stages of our method for editing a video with 200 frames on NVIDIA GeForce RTX 3090 Ti

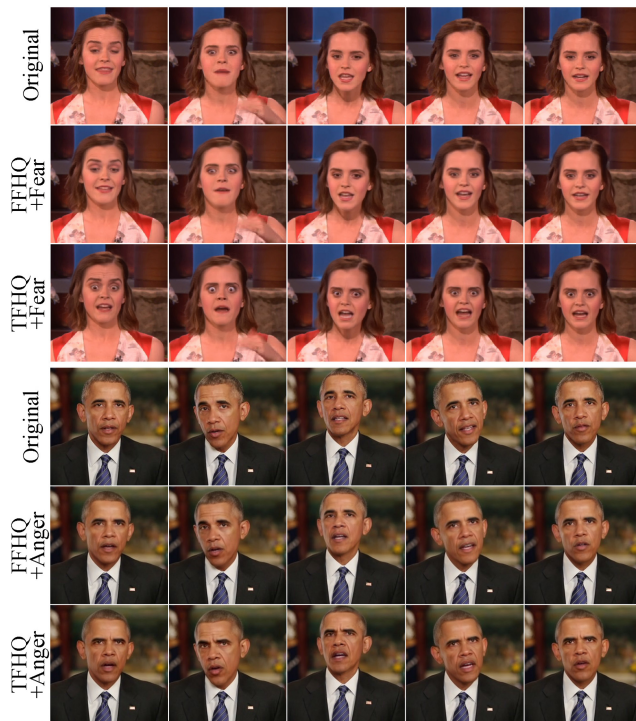
| Stage                                | Time   |
|--------------------------------------|--------|
| Encoder-based PTI                    | 15 min |
| Adaptive emotion amplification       | 20 ms  |
| Temporally-coherent stitching tuning | 19 min |

direction (e.g., fear and anger). The results in Fig. 15 clearly shows that our method based on StyleGAN-TFHQ can effectively perform amplification along the underrepresented emotion directions in StyleGAN, whereas the amplification by Baseline-2 along these directions is quite subtle.

#### 4.2.3 Quantitative results

We also quantitatively measured the quality of emotion amplification using our method and the baseline methods. We expect that high-quality emotion amplification results should edit facial expressions such that the correct prototypical emotions (*happiness, sadness, surprise, fear, anger, disgust, and contempt*) are more prominent in the resultant video frame. We also expected facial identity to remain consistent after editing. To validate this, we propose the following new metrics.

- EM-Con. This metric measures the prototypical emotion consistency between the original video frame and the resultant video frame. It is defined as the percentage of frames where the prototypical emotion is unchanged. We employ an off-the-shelf emotion classifier [54] to estimate the prototypical emotion based on the classification probability.



**Fig. 15** Our method based on StyleGAN-TFHQ results in more expressive emotion editing directions compared with the original StyleGAN trained on FFHQ. The emotions under-represented in FFHQ such as fear and anger can be more effectively amplified.

We expect that the editing process can magnify prototypical emotions in the right direction but not introduce other biased emotions.

- EM-New. This metric evaluates how many prototypical emotions are introduced from the neutral video frames after editing. We expect more non-neutral frames to be generated after emotion amplification.
- EM-Ratio. It measures how effectively the emotion is amplified by computing the ratio of the prototypical emotion probabilities before and after editing. We expect that the higher the ratio, the more effective the amplification is.
- ID-Sim: This metric evaluates the identity consistency before and after amplification based on an off-the-shelf facial identity detection network [56]. A higher score leads to better identity preservation.

Table 2 presents quantitative comparisons between the proposed and baseline methods. Our method is consistently better across all metrics compared with baseline methods. This confirms that our method can effectively magnify emotions, as expected. Baseline-1 has the lowest EM-Con and EM-Ratio because editing is performed in the middle layers of the  $W+$  space. While some amplification effects are noticeable in the results, they are not well disentangled from other attributes. This lack of disentanglement can result in unexpected changes and distortions, although it may incidentally make more faces appear non-neutral (see EM-New). Baseline-2 achieved more accurate results than Baseline-1 according to EM-Con and EM-Ratio. However, because of the use of an emotion-biased latent space, the amplification was not as effective as ours, given the worse scores under all metrics.

#### 4.2.4 Stitching-tuning method

We also quantitatively validated the effectiveness of our temporally coherent stitching-tuning technique. We measured the temporal coherence between two neighboring frames based on the flow warping error as [57, 58]:

**Table 2** Quantitative comparisons between our method and two baseline methods

| Method     | EM-Con $\uparrow$ | EM-New $\uparrow$ | EM-Ratio $\uparrow$ | ID-Sim $\uparrow$ |
|------------|-------------------|-------------------|---------------------|-------------------|
| Baseline-1 | 53.6%             | 48.3%             | 0.776               | 0.6381            |
| Baseline-2 | 66.0%             | 43.5%             | 0.959               | 0.8803            |
| Ours       | <b>75.4%</b>      | <b>51.3%</b>      | <b>1.063</b>        | <b>0.8892</b>     |

$$E_{\text{warp}}(I_i, I_{i+1}) = \frac{1}{\sum_{p=1}^N M_i^{(p)}} \sum_{p=1}^N M_i^{(p)} \|I_i^{(p)} - \hat{I}_{i+1}^{(p)}\|_1$$

where  $M_i \in [0, 1]$  is the mask for stitching the edited face,  $N$  is the number of pixels,  $M_i^{(p)}$  represents the  $p$ th pixel of  $M_i$ , and  $\hat{I}_{i+1}$  is the frame  $I_{i+1}$  warped by the optical flow  $F_{i \Rightarrow i+1}$ . The temporal coherence of the resulting video was defined as the average error of all the frames. We also measured relative temporal coherence with respect to the original video. Table 3 shows that our method results in better temporal coherence than the previous per-frame-based technique.

We further measured the sharpness preservation of the edited results using the cumulative probability of blur detection (CPBD) [59]. A higher CPBD score indicates fewer blurs detected, that is higher sharpness. The final score was calculated by dividing the CPBD score of each edited frame by that of the original frame. Table 4 shows the scores of STIT’s tuning method, the approach of tuning all frames simultaneously, and our approach. Our approach remarkably surpasses the all-frame approach and is slightly better than STIT.

### 4.3 Discussion

While our emotion amplification method works well on a range of facial videos with different prototypical emotions, it does not resolve the inherent limitations shared with existing works due to using the same StyleGAN2-based face editing approach, such as the “texture sticking” effect. The development of StyleGAN3 and its inversion techniques [60] would provide opportunities to further improve the results. Furthermore, since we adopt the same pipeline as in Ref. [6], challenging scenarios may cause

faces in shadows and/or under occlusions to be inaccurately cropped from the original video frame, resulting in unexpected editing outcomes. In our implementation, we excluded such frames from the evaluations; however, manually selecting the correct facial regions for amplification would benefit downstream applications in practice. In addition, the evaluation of emotion amplification is challenging because no ground truth is available for comparison. We proposed several new emotion-related quantitative metrics based on off-the-shelf emotion analysis tools. A better evaluation of these results requires further exploration.

## 5 Conclusions

In this study, we addressed the new problem of naturally amplifying the emotions of a facial video sequence. Unlike previous studies, where editing was often performed along a fixed semantic editing direction, our amplification was adaptively performed in the StyleGAN latent space. To ensure the quality of emotion amplification while preserving other attributes, we propose a carefully designed method that specifically edits faces in an emotion-editing subspace. This is achieved by extracting more effective editing directions associated with prototypical emotions through a fine-tuned StyleGAN whose latent space is less emotion-biased. Fine-tuning was based on a more emotion-balanced dataset TFHQ collected from online movie trailers with more diverse and expressive emotions from actors. Our proposed stitching tuning technique further ensures the temporal coherence of the resultant video. Both qualitative and quantitative evaluations using real video sequences demonstrated the capability of our method and its potential application in videos postproduction. In the future, we hope to make further improvements suited to natural and realistic video editing rather than applying methods for still images to video sequences.

## Acknowledgements

This study was supported by the National Key R&D Program of China (No. 2022YFF0902302), National Natural Science Foundation of China (No. 62322209), RCUK grant CAMERA (Nos. EP/M023281/1, EP/T022523/1), and a gift from Adobe.

**Table 3** Quantitative comparisons on temporal coherency between our stitching tuning method and STIT’s stitching tuning method

| Method | Warping error ↓ | Relative error ↓ |
|--------|-----------------|------------------|
| STIT   | 0.3533          | 100.83%          |
| Ours   | <b>0.3517</b>   | <b>100.21%</b>   |

**Table 4** Quantitative comparisons on sharpness preservation between STIT’s stitching tuning, tuning all frames simultaneously, and our stitching tuning method

| Method    | Sharpness preservation ↑ |
|-----------|--------------------------|
| STIT      | 91.56%                   |
| All-frame | 85.06%                   |
| Ours      | <b>92.94%</b>            |





### Availability of data and materials

The source code is publicly available at [https://github.com/ArgentRaider/Emotion\\_Amplification\\_StyleGAN\\_TFHQ](https://github.com/ArgentRaider/Emotion_Amplification_StyleGAN_TFHQ).

### Author contributions

Yukun Xu: theoretical development, experiment implementation, paper writing, approving the final version of the article publication, adding references. Justin N. M. Pinkney: theoretical development, experiment implementation, experimental design, paper writing, adding references. Yong-Liang Yang: theoretical development, experimental design, paper writing, approving the final version of the article publication, adding references. Tianjia Shao: theoretical development, experimental design, paper writing, approving the final version of the article publication, adding references. Kun Zhou: theoretical development, experimental design, paper writing, approving the final version of the article publication, adding references.

### Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article. The author Kun Zhou is the Associate Editor of this journal.

### Electronic Supplementary Material

Supplementary material is available in the online version of this article at <https://doi.org/10.26599/CVM.2025.9450391>.

### References

- [1] Karras, T.; Laine, S.; Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4396–4405, 2019.
- [2] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8107–8116, 2020.
- [3] Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-free generative adversarial networks. In: Proceedings of the 35th International Conference on Neural Information Processing Systems, Article No. 66, 2021.
- [4] Alaluf, Y.; Tov, O.; Mokady, R.; Gal, R.; Bermano, A. HyperStyle: StyleGAN inversion with HyperNetworks for real image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18490–18500, 2022.
- [5] Roich, D.; Mokady, R.; Bermano, A. H.; Cohen-Or, D. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics* Vol. 42, No. 1, Article No. 6, 2023.
- [6] Tzaban, R.; Mokady, R.; Gal, R.; Bermano, A.; Cohen-Or, D. Stitch it in time: GAN-based facial editing of real videos. In: Proceedings of the SIGGRAPH Asia Conference Papers, Article No. 29, 2022.
- [7] Shen, Y.; Gu, J.; Tang, X.; Zhou, B. Interpreting the latent space of GANs for semantic face editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9240–9249, 2020.
- [8] Denton, E.; Hutchinson, B.; Mitchell, M.; Gebru, T. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint* arXiv:1906.06439, 2019.
- [9] Abdal, R.; Zhu, P.; Mitra, N. J.; Wonka, P. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics* Vol. 40, No. 3, Article No. 21, 2021.
- [10] Härkönen, E.; Hertzmann, A.; Lehtinen, J.; Paris, S. GANSpace: Discovering interpretable GAN controls. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, Article No. 825, 2020.
- [11] Shen, Y.; Zhou, B. Closed-form factorization of latent semantics in GANs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1532–1540, 2021.
- [12] Jahanian, A.; Chai, L.; Isola, P. On the “steerability” of generative adversarial networks. *arXiv preprint* arXiv:1907.07171, 2019.
- [13] Bounareli, S.; Argyriou, V.; Tzimiropoulos, G. Finding directions in GAN’s latent space for neural face reenactment. *arXiv preprint* arXiv:2202.00046, 2022.
- [14] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. *arXiv preprint* arXiv:2103.00020, 2021.



- [15] Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; Lischinski, D. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2065–2074, 2021.
- [16] Abdal, R.; Zhu, P.; Femiani, J.; Mitra, N.; Wonka, P. CLIP2StyleGAN: Unsupervised extraction of StyleGAN edit directions. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings, Article No. 48, 2022.
- [17] Kocaszar, U.; Dirik, A.; Tiftikci, M.; Yanardag, P. StyleMC: Multi-channel based fast text-guided image generation and manipulation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 3441–3450, 2022.
- [18] Xu, Z.; Lin, T.; Tang, H.; Li, F.; He, D.; Sebe, N.; Timofte, R.; van Gool, L.; Ding, E. Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18208–18217, 2022.
- [19] Gabbay, A.; Cohen, N.; Hoshen, Y. An image is worth more than a thousand words: Towards disentanglement in the wild. In: Proceedings of the Conference on Neural Information Processing Systems, 9216–9228, 2021.
- [20] Tzelepis, C.; Oldfield, J.; Tzimiropoulos, G.; Patras, I. ContraCLIP: Interpretable GAN generation driven by pairs of contrasting sentences. *arXiv preprint arXiv:2206.02104*, 2022.
- [21] Abdal, R.; Qin, Y.; Wonka, P. Image2StyleGAN: How to embed images into the StyleGAN latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4431–4440, 2019.
- [22] Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; Cohen-Or, D. Encoding in style: A StyleGAN encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2287–2296, 2021.
- [23] Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; Cohen-Or, D. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics* Vol. 40, No. 4, Article No. 133, 2021.
- [24] Logacheva, E.; Suvorov, R.; Khomenko, O.; Mashikhin, A.; Lempitsky, V. DeepLandscape: Adversarial modeling of landscape videos. In: *Computer Vision—ECCV 2020. Lecture Notes in Computer Science*, Vol. 12368. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J.-M. Eds. Springer Cham, 256–272, 2020.
- [25] Wu, Z.; Lischinski, D.; Shechtman, E. StyleSpace analysis: Disentangled controls for StyleGAN image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12863–12872, 2021.
- [26] Tian, Y.; Ren, J.; Chai, M.; Olszewski, K.; Peng, X.; Metaxas, D. N.; Tulyakov, S. A good image generator is what you need for high-resolution video synthesis. *arXiv preprint arXiv:2104.15069*, 2021.
- [27] Fox, G.; Tewari, A.; Elgharib, M. A.; Theobalt, C. StyleVideoGAN: A temporal generative model using a pretrained StyleGAN. *arXiv preprint arXiv:2104.15069*, 2020.
- [28] Skorokhodov, I.; Tulyakov, S.; Elhoseiny, M. StyleGAN-V: A continuous video generator with the price, image quality and perks of StyleGAN2. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3616–3626, 2022.
- [29] Xu, Z.; Zhang, J.; Liew, J. H.; Zhang, W.; Bai, S.; Feng, J.; Shou, M. Z. PV3D: A 3D generative model for portrait video generation. *arXiv preprint arXiv:2212.06384*, 2022.
- [30] Oorloff, T.; Yacoob, Y. Expressive talking head video encoding in StyleGAN2 latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2998–3007, 2023.
- [31] Yao, X.; Newson, A.; Gousseau, Y.; Hellier, P. A latent transformer for disentangled face editing in images and videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 13769–13778, 2021.
- [32] Jiang, Y. R.; Chen, S. Y.; Fu, H.; Gao, L. Identity-aware and shape-aware propagation of face editing in videos. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 7, 3444–3456, 2024.
- [33] Zhao, S.; Wang, S.; Soleymani, M.; Joshi, D.; Ji, Q. Affective computing for large-scale heterogeneous multimedia data: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications* Vol. 15, No. 3s, Article No. 93, 2019.
- [34] Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* Vol. 13, No. 3, 1195–1215, 2022.
- [35] Tian, Y. I.; Kanade, T.; Cohn, J. F. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 23, No. 2, 97–115, 2001.
- [36] Feng, X.; Pietikäinen, M.; Hadid, A. Facial expression

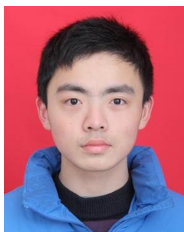


- recognition based on local binary patterns. *Pattern Recognition and Image Analysis* Vol. 17, No. 4, 592–598, 2007.
- [37] Shan, C.; Gong, S.; McOwan, P. W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing* Vol. 27, No. 6, 803–816, 2009.
- [38] Dahmane, M.; Meunier, J. Emotion recognition using dynamic grid-based HoG features. In: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, 884–888, 2011.
- [39] Lorincz, A.; Jeni, L. A.; Szabó, Z.; Cohn, J. F.; Kanade, T. Emotional expression classification using time-series kernels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 889–895, 2013.
- [40] Goodfellow, I.; Erhan, D.; Carrier, P.; Courville, A. C.; Mirza, M.; Hamner, B.; Cukierski, W. J.; Tang, Y.; Thaler, D.; Lee, D. H.; et al. Challenges in representation learning: A report on three machine learning contests. In: *Neural Information Processing. Lecture Notes in Computer Science, Vol. 8228*. Lee, M.; Hirose, A.; Hou, Z.-G.; Kil, R. M. Eds. Springer Cham, 117–124, 2013.
- [41] Benitez-Quiroz, C. F.; Srinivasan, R.; Martinez, A. M. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5562–5570, 2016.
- [42] Mollahosseini, A.; Hasani, B.; Mahoor, M. H. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* Vol. 10, No. 1, 18–31, 2019.
- [43] Yu, Z.; Zhang, C. Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the ACM on International Conference on Multimodal Interaction, 435–442, 2015.
- [44] Ebrahimi Kahou, S.; Michalski, V.; Konda, K.; Memisevic, R.; Pal, C. Recurrent neural networks for emotion recognition in video. In: Proceedings of the ACM on International Conference on Multimodal Interaction, 467–474, 2015.
- [45] Hasani, B.; Mahoor, M. H. Facial expression recognition using enhanced deep 3D convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2278–2288, 2017.
- [46] Yang, H.; Ciftci, U.; Yin, L. Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2168–2177, 2018.
- [47] Kumawat, S.; Verma, M.; Raman, S. LBVCNN: Local binary volume convolutional neural network for facial expression recognition from image sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 207–216, 2019.
- [48] Daněš, R.; Black, M.; Bolkart, T. EMOCA: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 20279–20290, 2022.
- [49] Blanz, V.; Vetter, T. A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, 187–194, 1999.
- [50] Haas, R.; Graßhof, S.; Brandt, S. S. Tensor-based emotion editing in the StyleGAN latent space. *arXiv preprint arXiv:2205.06102*, 2022.
- [51] Qi, D.; Tan, W.; Yao, Q.; Liu, J. YOLO5Face: Why reinventing a face detector. In: *Computer Vision – ECCV 2022. Lecture Notes in Computer Science, Vol. 13805*. Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; Hassner, T. Eds. Springer Cham, 228–244, 2023.
- [52] Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2018.
- [53] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.
- [54] Cheong, J. H.; Jolly, E.; Xie, T.; Byrne, S.; Kenney, M.; Chang, L. J. Py-feat: Python facial expression analysis toolbox. *arXiv preprint arXiv:2104.03509*, 2021.
- [55] Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; Cohen-Or, D. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics* Vol. 40, No. 4, Article No. 133, 2021.
- [56] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. ArcFace: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4685–4694, 2019.
- [57] Lai, W.-S.; Huang, J.-B.; Wang, O.; Shechtman, E.; Yumer, E.; Yang, M.-H. Learning blind video temporal consistency. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11219*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 179–195, 2018.





- [58] Lei, C.; Xing, Y.; Chen, Q. Blind video temporal consistency via deep video prior. In: Proceedings of the 34th Conference on Neural Information Processing Systems, 1083–1093, 2020.
- [59] Narvekar, N. D.; Karam, L. J. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Transactions on Image Processing* Vol. 20, No. 9, 2678–2683, 2011.
- [60] Alaluf, Y.; Patashnik, O.; Wu, Z.; Zamir, A.; Shechtman, E.; Lischinski, D.; Cohen-Or, D. Third time's the charm? Image and video editing with StyleGAN3. *arXiv preprint arXiv:2201.13433*, 2022.



**Yukun Xu** is a postgraduate at the State Key Laboratory of CAD&CG, Zhejiang University. His research interests include computer graphics and digital human subjects.

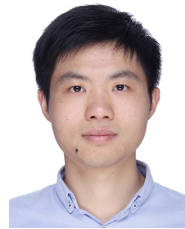


**Justin N. M. Pinkney** worked as a senior machine learning researcher at Lambda, Inc. He received his Ph.D. degree in biophysics at the University of Oxford. He focused on computer vision and generative models.



**Yong-Liang Yang** is a senior lecturer in the Department of Computer Science at the University of Bath, UK. He received his B.Sc. and Ph.D. degrees in computer science from Tsinghua University. His research area is in visual computing, with particular interest in shape modelling, computational design, and interactive

techniques.



**Tianjia Shao** is a professor in the State Key Laboratory of CAD&CG, Zhejiang University. He was previously a lecturer at the School of Computing, University of Leeds, UK. He received his Ph.D. degree in computer science from the Institute for Advanced Study, Tsinghua University, and his B.Sc. degree from the Department of Automation, Tsinghua University. His research focuses on 3D modeling, digital humans, and computer animation.



**Kun Zhou** is a Cheung Kong Professor in the Computer Science Department of Zhejiang University, and the Director of the State Key Lab of CAD&CG. Prior to joining Zhejiang University in 2008, Dr. Zhou was a leading researcher at the Internet Graphics Group at Microsoft Research Asia. He received his B.Sc. and Ph.D. degrees in computer science from Zhejiang University in 1997 and 2002, respectively. His research interests include visual computing, parallel computing, human–computer interaction, and virtual reality. He is a Fellow of the IEEE and ACM.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

To submit a manuscript, please go to <https://jcvn.org>.



清华大学出版社  
Tsinghua University Press

Available on  
**IEEE Xplore®**