A StyleCLIP-based Facial Emotion Manipulation Method for Discrepant Emotion Transitions

Qi Guo[1][0000−0002−6076−6126] and Xiaodong Gu[1][0000−0002−7096−1830]

Department of Electronic Engineering, Fudan University,Shanghai 200438,China
xdgu@fudan.edu.cn

Abstract. Leveraging StyleCLIP's expressivity and its disentangled latent codes, current methodologies enable facial emotion manipulation through textual inputs. Despite these advancements, significant challenges remain in manipulating target emotions that deviate markedly from the originals, especially in transitioning from happy to sad emotions without introducing artifacts or errors. This paper introduces a novel approach for discrepant emotion transitions. Our network architecture integrates a StyleGAN2 generator with an Emotion Manipulation Mapper, a Dual Auxiliary Classifier, and a CLIP Text Encoder. By utilizing the inverse cumulative distribution function, we convert source emotion labels into conditional data, thus enhancing the model's ability to accurately map and modify the emotional distribution across faces. We evaluated our method against established techniques using the Radboud Faces Database and CelebA-HQ dataset, and introduced a new quantitative measure including seven metric for assessing manipulation efficacy.

Keywords: Multimodality · Facial emotion manipulation · StyleCLIP.

1 Introduction

Facial emotions such as anger, disgust, fear, happiness, sadness, and surprise directly reflect psychological states. Emotion manipulation technology holds immense value across various fields, providing innovative tools for academic research and broad application prospects in medical care, entertainment, virtual reality, and human-computer interaction. As technology [2, 4, 52, 53] continually progresses, the widespread application of emotion manipulation is anticipated to grow, offering significant social and economic benefits. This highlights the essential demand for advanced tools capable of accurately transforming source emotions into specified target emotions while maintaining an individual's facial identity.

However, existing methods [54, 55, 35, 44] often introduce unnecessary or irrelevant details during the emotional manipulation process, potentially disrupting the intended outcomes. For instance, in virtual communication, imprecise facial emotion manipulation tools can allow users to incorrectly convey their desired emotions, increasing misunderstandings in relationships. Although there

has been significant progress in improving the accuracy of facial emotion manipulation [6, 17, 16, 20], notable challenges persist, particularly when modifying emotions that drastically differ from the original state, such as transitioning from happiness to sadness. Certain methods [3, 17, 57, 20, 21] successfully manipulate emotions like anger and disgust, which are similar, as indicated by a "✓ " mark in Figure 1. However, these methods struggle with transitions involving significant emotional differences, such as from anger to happiness, marked by an "× " . Thus, improving facial emotion manipulation techniques to enhance stability and expand the range of compatible input emotions is critical for meeting the demands of digital-age remote socialization, enabling more intricate and empathetic virtual interactions.

In sum, current methods have shown some efficacy in emotional manipulation, they often fall short when managing transformations involving large emotional disparities, highlighting the necessity for further enhancements. This paper introduces an efficient strategy to transform complex, non-neutral emotions into specific target emotions, ensuring precise alterations in emotion distribution and enabling seamless one-to-any emotion manipulation.

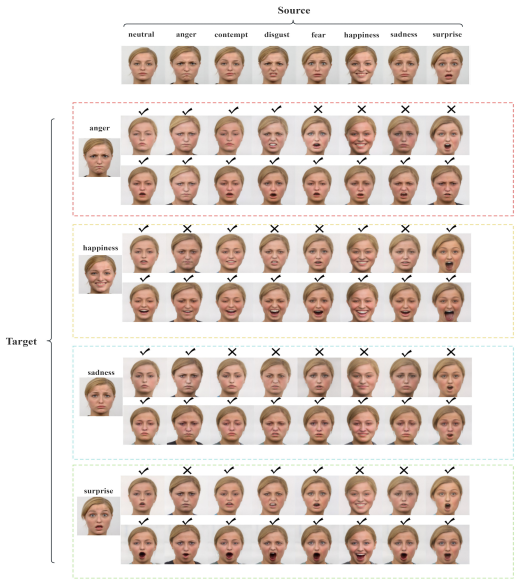

Fig. 1. Comparative outcomes of one-to-any facial emotion manipulation across various methods. This visualization illustrates that when there is a substantial discrepancy between the target and source emotions, the outcomes are generally unsatisfactory, as denoted by the "× " mark.

2 Related Works

2.1 Conditional Face Image Manipulation

Building on the advancements in text-guided face image manipulation, the field of conditional face image manipulation has similarly advanced, offering enhanced control over the direction of image transformations. This evolution emphasizes the need for improved efficiency, flexibility, and controllability in image manipulation techniques. A notable development in this area is StyleMC [44], which adopts a text as condition for image generation and manipulation. This method has demonstrated that images can be accurately manipulated with a single text prompt without significantly altering other attributes, presenting a faster and more efficient alternative to earlier techniques that required extensive data volumes or complex pre-processing. TediGAN [43] extends these capabilities further by supporting multi-modal image generation and manipulation. This advancement marks a considerable progression in the field, as it accommodates various inputs while ensuring high-quality outputs, paving the way for more diverse and adaptable face manipulation methods. Recent studies, such as the one presented in [10], propose pretraining an attribute prediction model by inverting synthesized face images back into the GAN latent space. This technique specifically addresses the semantics encoded within the latent space of a pretrained GAN, enhancing the relevance and accuracy of generated images. FaceComposer [11] represents another innovative approach within the latent diffusion framework. It adheres to the compositional generation paradigm, utilizing face-specific conditions such as Identity Feature and Projected Normalized Coordinate Code to maximize creative output from the model. However, this method requires substantial training resources, highlighting a significant challenge in scalability and resource allocation.

While the addition of conditions to these models is widely recognized as beneficial, there remains a considerable need for further research into how these conditions can be integrated more effectively. Future studies should focus on selecting the most relevant conditions and refining how these conditions are articulated to make the direction of manipulation more precise. This ongoing refinement is crucial as the field continues to strive for methods that offer more precise control and greater flexibility in face image manipulation.

2.2 Facial Emotion Manipulation

Facial emotion manipulation represents an advanced tier of face manipulation, demanding a nuanced understanding of facial dynamics and a deep knowledge of human emotion. This process strives to accurately transition from a source emotion to a target emotion while meticulously preserving other facial attributes, which has significant implications across various applications. GANs have emerged as a robust tool in this domain, facilitating more sophisticated manipulations of facial emotion. For instance, StarGAN [27] exemplifies a versatile approach capable of translating images across multiple domains using a

single model, which simultaneously preserves identity features by minimizing cycle loss. However, StarGAN is generally restricted to generating discrete emotion and does not offer the flexibility required for creating continuous or subtle changes in facial emotion. To overcome some of these limitations, a novel interdisciplinary method introduced by Yan et al. [14] merges Generative Adversarial techniques with insights from cognitive sciences, specifically employing psychophysical reverse correlation. This method is a data-driven approach that effectively captures an observer's mental representation of a specific facial expression, enhancing the precision of emotion manipulation. Additionally, advancements in multimodal approaches have seen developments like the Sound-to-Expression (S2E) framework [13], which leverages emotional cues from sounds to guide the generation of facial emotion, supported by a specially constructed speech dataset for emotion recognition. Research in 3D facial manipulation also continues to evolve. Techniques leveraging 3D Morphable Models (3DMM) combined with StyleGAN texture maps present a comprehensive strategy for capturing intricate facial movements and appearance details, embedding emotional information effectively. However, such methods [12] require substantial computational resources and are subject to further refinement to eliminate errors or undesirable effects.

As the field of facial emotion manipulation evolves, the integration of diverse techniques and the enhancement of existing methods remain pivotal for advancing the precision and applicability of facial emotion manipulation, underscoring the ongoing need for research that addresses the technical challenges.

3 Methodology

To achieve satisfactory manipulation outcomes, even when dealing with substantial discrepancies between original and target emotions, we introduce a sophisticated approach, as illustrated in Figre. 2. Our method incorporates an Emotion Manipulation Mapper, a Dual Auxiliary Classifier, a CLIP text encoder, and a pre-trained StyleGAN2 generator. To enhance the accuracy of manipulating complex, non-neutral emotions and to ensure that the identity of the subject is preserved, our tailored loss function, $\mathcal{L}_{emotion}$, simultaneously considers both the accuracy of the emotional manipulation and the preservation of the individual's identity. This holistic approach aims to transform complex, non-neutral emotions into specific target emotions, ensuring precise alterations in emotion distribution and enabling seamless one-to-any emotion manipulation.

3.1 Emotion Manipulation Mapper

To address the challenges posed by significant disparities between the input and target emotion distributions in facial emotion manipulation, this paper presents a refined approach. The method aims to enhance the model's capacity to accurately pinpoint the correct manipulation direction by enhancing its domain awareness. This is crucial for achieving high fidelity in the transition from the

source to the target emotion while preserving other facial attributes, a process fundamental in broad applications ranging from entertainment to therapeutic contexts.

We introduce the Emotion Manipulation Mapper (EMM), depicted in Figure 2, which integrates supplementary conditional information generated by the Inverse Cumulative Distribution Function (ICDF). This integration allows the model to evaluate the original input emotion effectively, making it more sensitive to the nuances between different emotional distributions. This sensitivity is critical for successfully executing facial emotion manipulation in the $W+$ space with any provided latent code $w \in W+$.

The latent code of the input image, denoted as widwid, captures essential facial attributes such as hairstyle and face shape. To enhance the model's adaptability to various emotions, the ICDF [50] is employed. This function translates an emotion label into a conditional embedding constrained by a predefined noise distribution. Specifically, the ICDF determines the value corresponding to a specific emotion's cumulative probability, effectively mapping the standard normal distribution input into a uniform distribution which is then transformed back to target emotion values.

Given a data row R with $R = [r_1, r_2, \ldots, r_8]$, each r_i is a random sample from a standard normal distribution, symbolizing a distinct emotion. Each r_i is transformed into a uniformly distributed data point via the Cumulative Distribution Function (CDF), which is defined as Equation 1:

$$F(x) = P(X \leq x), \tag{1}$$

where X is a random variable from a normal distribution. $P(X \leqslant x)$ represents the probability that the random variable X is less than or equal to x. Applying this to each r_i yields as Equation 2:

$$u_i = F(r_i), \tag{2}$$

where each u_i residing in the interval [0,1].

For each emotion, a distinct ICDF, represented as F^{-1}, is established. Utilizing u_i, the original data value is ascertained as Equation 3:

$$v_i = F^{-1}(u_i), \tag{3}$$

where v_i denotes the value derived from the ICDF evaluation of u_i. Each v_i subsequently acts as the condition embedding for its respective emotion, denoted $w_{condition}$.

Subsequently, w_{id} and $w_{condition}$ are amalgamated as Equation4 to derive w_{fused} Equation 4:

$$w_{fused} = w_{id} \oplus w_{condition}. \tag{4}$$

The resultant vector, w_{fused}, is then input into the Emotion Editor (EE). Although the architecture of EE is inspired by StyleGAN's mapping network,

it employs a truncated layer set, comprising four layers instead of the conventional eight. EE further refines $w_{emotion}$ for a specific text prompt to deduce a manipulation step as Equation 5:

$$w_{emotion} = EE(w_{fused}), \tag{5}$$

leading to the final augmentation of the original w as Equation 6:

$$w_{EMM} = w \oplus w_{emotion}, \tag{6}$$

where $\oplus$ represents the concatenation operation. The concatenated vector w_{EMM} is then input into both the Generator and the Dual Auxiliary Classifier.

In conclusion, the primary goal of the Emotion Manipulation Mapping EMM network is to enhance the model's domain awareness. This enables it to adeptly adjust to the distributional disparities among various emotions through the integration of additional conditions, managed by the ICDF. EMM proves particularly efficient in cases where there is a pronounced disparity between the source emotional distribution and the target emotional outcome. Without such an operation, the model might struggle to recognize and adapt to these distributional nuances. By employing the Emotion Manipulation Mapping, we gain deeper insight into the mechanics of emotion manipulation, enabling a more accurate expression of emotions.

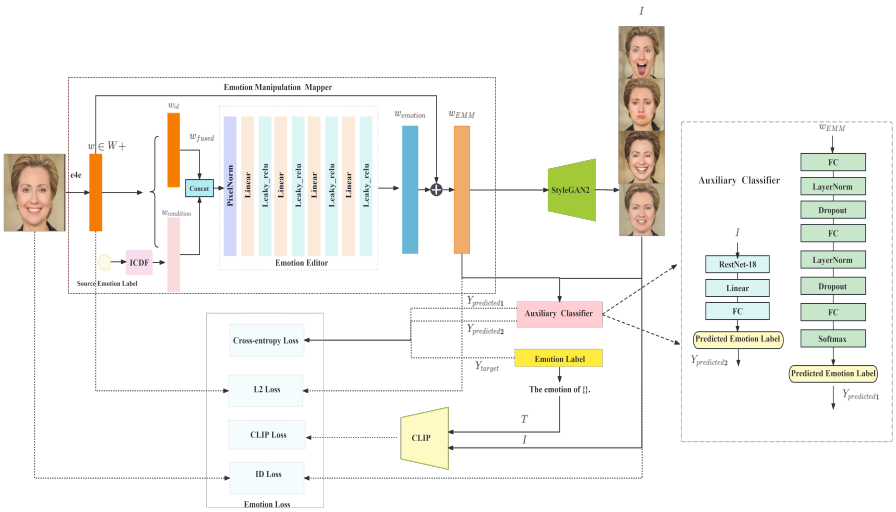

Fig. 2. The framework of the proposed methods. Our method comprises an Emotion Manipulation Mapper, a Dual Auxiliary Classifier, a CLIP text encoder, and a pretrained StyleGAN2 generator.To increase the manipulation accuracy of non-neutral complex emotions and maintain identity preservation, our loss function $\mathcal{L}_{emotion}$ takes into account both manipulated accuracy and identity preservation.

3.2 Dual Auxiliary Classifier

To further refine the accuracy and robustness of our model in facial emotion manipulation, we have integrated a Dual Auxiliary Classifier (DAC) mechanism. This innovative approach enhances the model's capabilities by providing multiple sources of label information, thus guiding the manipulation process more effectively.

The Dual Auxiliary Classifier operates on two levels. it analyzes both the latent code and the generated image simultaneously. Specifically, the latent code w_{EMM} is assessed by a pre-trained classifier designed for latent codes, while the generated image I_m is evaluated by a pre-trained image classifier. This dual approach ensures a comprehensive analysis of both the input data and the manipulated output, leading to the prediction of emotions denoted as $Y_{predicted1}$ and $Y_{predicted2}$, respectively.

In our DAC design, we utilize two separate sub-classifiers. Both are trained on 2,000 images from the Affectnet dataset alongside their corresponding labels [51]. Our latent code classifier is designed to accept latent codes $w \in \mathbb{R}^{18\times512}$. The e4e method [46] is employed to transform images into latent codes, aligning with our model's use of the CelebA-HQ dataset [49] for training. The architecture of the latent code classifier can be expressed as Equation 7:

$$Y_{predicted1} = f_{\text{Dropout}}\left(f_{\text{LN2}}\left(f_{\text{FC2}}\left(f_{\text{Dropout}}\left(f_{\text{LN}}\left(f_{\text{FC}}(w_{\text{EMM}})\right)\right)\right)\right)\right). \tag{7}$$

Concurrently, the image classifier accepts image $I_m \in \mathbb{R}^{3\times512\times512}$. We implement a pre-trained Resnet-18 network, and its architecture can be captured as Equation 8:

$$Y_{predicted2} = f_{\text{FC}}(f_{\text{Linear}}(f_{\text{Resnet-18}}(I_m))). \tag{8}$$

Both classifiers output emotion labels, each systematically mapped from 0 to 7. This numerical delineation of emotions facilitates the machine's comprehension and offers a quantifiable scale for emotions. The cross-entropy loss function between the predicted and target emotion labels is computed to fine-tune the manipulation accuracy.

In conclusion, the integration of the Dual Auxiliary Classifier into our model represents a significant advancement in text-driven facial emotion manipulation. By leveraging rich label data and sophisticated classification techniques, our approach significantly enhances the precision and effectiveness of the manipulation outcomes. This dual classifier system not only improves model performance but also pushes the boundaries of what is possible in the field of emotion manipulation, providing a robust framework for future enhancements.

3.3 Emotion Loss Function

To further refine the manipulation direction of non-neutral complex emotions and enhance manipulation accuracy, our comprehensive loss function, $\mathcal{L}_{emotion}$,

is divided into two primary components: manipulated accuracy and identity preservation. This structured approach ensures that each aspect of the facial emotion manipulation process is optimally controlled, enhancing the overall effectiveness and fidelity of the generated images.

The first task is to measure manipulated accuracy, which meets the target text description and the authenticity of emotions. $\mathcal{L}_1$ includes the following loss functions.

CLIP Loss: Building upon the methodologies of previous works [36, 16, 37], we utilize a directional CLIP loss. The CLIP loss $\mathcal{L}_{\text{CLIP}}\ (I)$ guides the Emotion Manipulation Mapper to minimize the cosine distance in the CLIP latent space as Equation 9:

$$\mathcal{L}_{\text{CLIP}}\ (I, T) = D_{\text{CLIP}}\ \ (I_m, T)\,, \tag{9}$$

where D_{CLIP} represents the cosine distance in facial emotion space. I_m is manipulated image generated by $G\,(w_{EMM})$, G refers to the pre-trained StyleGAN2 generator, T is the text obtained from the text prompt "an emotion of {}" with the corresponding emotion label, and EMM denotes the Emotion Mapper network.

Cross-entropy Loss: To effectively uses the capability of Dual Auxiliary Classifier in making accurate predictions to help guide the manipulation process of the objective function. In our model, we utilize the sigmoid function for the output of the network's final layer. Each neuron in this layer corresponds to an emotion category. The output from each neuron indicates the probability of the presence of its associated emotion. Given that our multi-emotion classification task treats each emotion category as independent, the sum of the probabilities from the final layer's neurons does not necessarily equal 1. For one emotion of tasks in multi-emotion classification tasks, the cross-entropy loss function as Equation 10:

$$\begin{aligned}\mathcal{L}_{\text{cross-entropy}}(w) = -\,[&P(Y_t)\cdot\log(Q(Y_p))+\\ &(1-Q(Y_p))\cdot\log(1-P(Y_t))]\,,\end{aligned} \tag{10}$$

where Y_t is the target label transformed from the given text. $P(Y_t)$ represents the true target label distribution for each emotion category, which follows a binomial distribution with potential values of 0 or 1. On the other hand, $Q(Y_p)$ denotes the network's predicted emotion label probability distribution. In our multi-emotion classification context, if we analyze each emotion category separately, we can interpret Q as the likelihood of the target label Y_t being present.

The loss of the first task is as Equation 11:

$$\begin{aligned}\mathcal{L}_1 = \lambda_{CLIP}\mathcal{L}_{\text{CLIP}}(I,T)+\\ \lambda_{\text{cross-entropy}}\mathcal{L}_{\text{cross-entropy}}(w).\end{aligned} \tag{11}$$

The second task of our model is to ensure identity preservation, which is vital for maintaining the realism and natural appearance of the manipulated image

while preserving the identity of the subject. The loss function $\mathcal{L}_2$ encompasses several specific loss components to achieve this goal.

$L2$ **Loss:** To preserve the visual attributes of the original input emotion image, we minimize the $L2$ norm of the manipulation step in the latent space as Equation 12:

$$\mathcal{L}_{L2} = \|(w - w_{EMM})\|_2 , \tag{12}$$

where $\|\|_2$ is the $L2$ distance.

Identity (ID) Loss: To retain the identity attributes of the source emotion image during the manipulating process, we introduce an identity (ID) loss. This loss ensures that facial features remain consistent, preventing unwanted changes in the subject's identity. The ID loss is formulated as Equation 13:

$$\mathcal{L}_{\text{ID}}(I) = 1 - \langle R(I_s), R(I_m) \rangle , \tag{13}$$

where R is a pretrained ArcFace network for face recognition. I_m is the manipulated images generated by $G(w_{EMM})$, I_s is source images generated by $G(w)$. The operator$\langle \cdot, \cdot \rangle$ calculates the cosine similarity between features extracted from the source and target images.

The loss of the second task is as Equation 14:

$$\mathcal{L}_2 = \lambda_{L2}\mathcal{L}_{L2}(w) + \lambda_{\text{ID}}\mathcal{L}_{\text{ID}}(I). \tag{14}$$

These components of $\mathcal{L}_2$ collectively ensure that the manipulated images are not only effective in terms of emotion alteration but also remain true to the original identity of the subject, thus enhancing the authenticity and applicability of the generated results.

Our total loss function is a weighted combination of these losses as Equation 15:

$$\mathcal{L}_{\text{emotion}}(w, I, T) = \mathcal{L}_1 + \mathcal{L}_2. \tag{15}$$

The parameter values we use for the examples in this paper are $\lambda_{\text{CLIP}} = 1, \lambda_{\text{cross-entropy}} = 0.1, \lambda_{\text{L2}} = 0.8, \lambda_{\text{ID}} = 0.5$.

4 Experiments

In this section, we evaluate the effectiveness of our proposed methods for facial emotion manipulation tasks and demonstrate their superiority over existing state-of-the-art approaches including StyleCLIP [35],TediGAN [43] and StyleMC [44]. Detailed results and comparisons are presented in the subsections that follow. All experiments were conducted using the PyTorch framework on an NVIDIA RTX 3080ti GPU. We select the entire CelebA-HQ dataset, comprising 30,000 celebrity portraits, as our training set. We perform zero-shot experiment on the whole Radboud Faces Database (RaFD).

4.1 Evaluation Metrics

To ensure a comprehensive evaluation of our proposed method for facial emotion manipulation, we assess it from three dimensions: manipulated image quality, identity preservation, and manipulation accuracy. These evaluations involve comparing manipulated images I_m, source images I_s, and target images I_t, as well as analyzing both the initial and manipulated latent codes of the images, denoted w and w_{EMM}, respectively. The target label is represented as Y_t and the predicted label as Y_p.

Our evaluation metrics are exceptionally comprehensive and allow for a nuanced quantitative analysis of the model's performance in facial emotion manipulation. This quantity evaluation not only assesses basic image manipulations but also evaluates how these manipulations affect perceived identity and emotion, providing a holistic view of the technology's capabilities and limitations. For manipulated image quality, we use Frechet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS). For identity preservation, we use the Identity Preservation (IP) metric, which evaluates whether the identity of the person in the image remains the same after manipulation. The Multi-Scale Structural Similarity Index Measure (MS-SSIM) assesses the richness of details in the manipulated images, with higher MS-SSIM values indicating greater perceptual similarity to the target images in terms of detail. For manipulation accuracy, Emotion Distance (ED) evaluates whether the desired emotion has been accurately applied to the manipulated image using an emotion classifier to predict the emotion in the manipulated image. Action Unit Scores (AUs) measures the accuracy of emotion manipulation, where a lower distance indicates better manipulation accuracy. Lastly, CLIP Scores (CS) evaluates the consistency between the manipulated image and the text description by computing the cosine similarity between the normalized image and text embeddings.

4.2 Qualitative Comparison

The visualization results show the outcomes of different methods for transitioning multiple non-neutral complex emotions to other emotions, as shown in Figure 3. Upon examining the visual results, it's clear that our method generates more realistic outcomes in terms of both emotion accuracy and the detailing of facial features. In comparison, existing methods often struggle to substantially change target emotions from their original state, resulting in distorted and troublesome outcomes during the manipulation process. Our approach excels in its robustness for accurately manipulating emotions while maintaining the character's identity. Not only is our method robust, but it also shines in preserving intricate facial attribute details.

In conclusion, our method exhibits strength and precision, and it stands out in preserving facial attributes and the character's identity during emotional adjustments. It consistently produces high-quality and accurate results. It's worth mentioning that even though we use the same pre-trained StyleGAN2 as the generator for all comparative methods, our approach still outperforms.

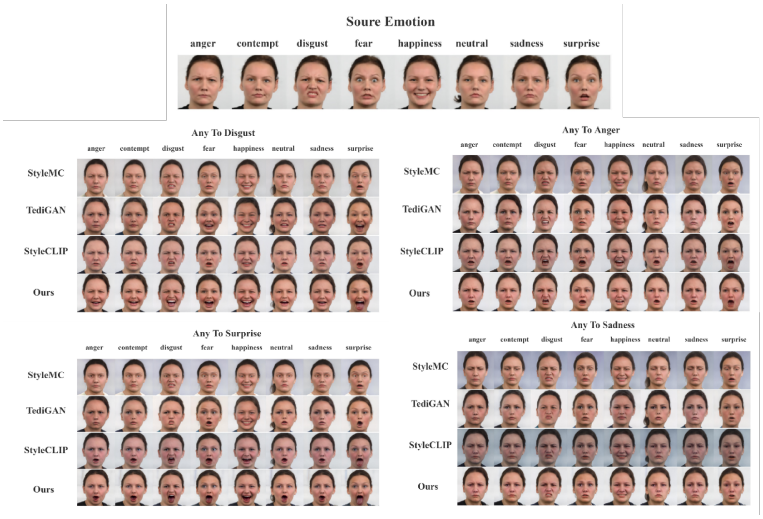

Fig. 3. The quality results of manipulating various emotions into anger using various methods. The visualization results show our method exhibits strength and precision, and it stands out in preserving facial attributes and the character's identity during emotional adjustments.

4.3 Quantitative Evaluation

Table 1. Quantitative comparisons with the state-of-the-art in RaFD dataset. m represents the manipulated images, s represents the source emotion images, and t represents the target emotion images. **Bold** indicates the best effect and underline indicates the second best effect.

Method	FID↓	LPIPS↓	MS-SSIM↑	IP↑	CS↑	Aus↓	ED↑
StyleMC	91.594	0.432	0.603	0.392	0.578	0.228	0.725
TediGAN	<u>87.601</u>	0.414	0.737	0.369	0.649	0.224	0.764
StyleCLIP	**86.858**	0.415	**0.778**	0.378	0.649	0.217	0.722
Ours	88.688	**0.288**	<u>0.755</u>	**0.425**	**0.655**	**0.216**	**0.830**

When evaluating our proposed method for facial emotion manipulation, we adopt a comprehensive approach, assessing the method's performance across various metrics crucial for practical implementations. The results of these comparisons, presented in Table 1, illustrate the superior performance of our method across the comprehensive analysis.

Firstly, compared to other methods, our model significantly outperforms in the LPIPS metric, although it does not achieve the best score in FID. LPIPS is more sensitive to local changes in images, such as subtle edits to facial features and detail enhancement, and can more accurately reflect the quality of the

edited image. In contrast, FID, being a global quality assessment metric, may not be sensitive enough to these local changes. This further illustrates the precise manipulation capabilities of our method in emotion manipulation.

Secondly, we use IP scores to compute the cosine similarity between the manipulated image and the target emotion images. The data clearly shows that our method excels in retaining facial identity. Additionally, our method stands out in preserving facial feature details, as indicated by the second-best MS-SSIM score . This score reveals that our outcomes are closely aligned with the facial details of the target emotion images, underscoring our method's ability to deliver both intricate detail and realism in emotion manipulation. This indicates that our method maintains a good balance in preserving facial identity while enhancing details.

Thirdly, considering that the methods we're comparing are all text-based, we begin by using CLIP scores to gauge the similarity between text and images. Our results unequivocally show that our method provides optimal alignment between text and images, successfully fulfilling our objective of text-driven optimization. When we employ a trained emotion classifier to discern the specific emotion, our method boasts the highest ED score, underscoring the precision of our emotion manipulation. It is well-established that varying facial feature movements play a crucial role in emotional expression. To rigorously assess our model's proficiency in manipulating facial emotion, we employ AUs to measure the activity levels across 17 distinct human facial emotions. The comparison reveals that the disparity in AUs between the images manipulated using our method and the target images is minimal, further attesting to our method's accuracy.

Our comprehensive set of metrics, focusing on the quality, realism, and accuracy of emotion manipulation, highlights our approach's exceptional capabilities. The model excels not only in replicating detailed emotional nuances but also in preserving the identity and intricate attributes of characters.

5 Ablation Study

In this ablation study, we dissect the individual contributions of specific components within our method, illustrating their impact through extensive testing as depicted in Figure 4. The quality scores are captured in Table 2, where m represents the manipulated images, s represents the source emotion images, and t represents the target emotion images.

5.1 Effect of Dual Auxiliary Classifier

In this section, we dissect the contributions of distinct components within our Dual Auxiliary Classifier (DAC) module.Table 2 shows how the DAC plays a key role in boosting manipulation accuracy over the baseline StyleCLIP. As Figure 4 points out, emotional manipulations can fall short when the DAC is left out. This highlights its importance in fine-tuning emotion transitions, particularly in facial areas like the eyes and mouth. The DAC provides a solid grip over

Fig. 4. The quality results of manipulating various emotions into happy and sad emotion using using various parts of our proposed method.

Table 2. Comparative analysis of different components of our proposed methods across various metrics. **Bold** indicates the best effect, and underline indicates the second best effect.

Method	FID↓	LPIPS↓	MS-SSIM↑	IP↑	CS↑	Aus↓	ED↑
StyleCLIP	86.858	0.415	0.778	0.378	0.649	0.217	0.722
StyleCLIP-DAC	99.225	0.431	0.784	0.378	0.669	**0.202**	0.748
LC-EMM-StyleCLIP-DAC	**80.612**	0.320	**0.788**	0.366	**0.677**	0.207	0.774
GC-EMM-StyleCLIP-DAC	88.688	**0.288**	0.755	**0.425**	0.655	0.216	**0.830**

the transition from non-neutral to target emotion distributions, enhancing the authenticity of subtle changes. The superior performance of the StyleCLIP-DAC method in terms of Aus scores confirms that DAC offers precise guidance in the direction of manipulation.

5.2 Effect of Emotion Manipulation Mapper

In this section, we'll explore how our Emotion Manipulation Mapper (EMM) module's components work in harmony to manipulate complex emotions effectively, maintaining detail in the results. This is particularly evident in challenging transitions such as shifting from any emotion to happiness or sadness, as depicted in Figure 4. To identify where to insert additional information processed by the Inverse Cumulative Distribution Function (ICDF) within the image latent code, we focused on the facial features with the most significant impact on emotional expression. We compared the local codes (w_{local}), which blend coarse and medium resolutions (known as LC-EMM-StyleCLIP), with the global codes (w_{global}) that span all resolutions (referred to as GC-EMM-StyleCLIP), the primary encoding in our approach. This comparison illuminated how adding $w_{condition}$ to these codes and then inputting them into the generator influences emotion manipulation. The metric results show that the LC-EMM-StyleCLIP-DAC method performs best in terms of FID, LPIPS, and MS-SSIM, whereas the GC-EMM-StyleCLIP-DAC method excels in IP and ED . However, it's crucial

to note that effective emotion manipulation involves more than just transferring emotions—it also preserves the original character's identity. Our findings demonstrate that combining $w_{condition}$ with w_{global} allows for both precise manipulation and the retention of character attributes, achieving a balance between detailed emotional expression and identity preservation.

6 Conclusion

This paper introduces a novel StyleCLIP-based domain-aware facial emotion manipulation method that integrates dual auxiliary classifiers to enhance the state-of-the-art in facial emotion synthesis. By leveraging StyleCLIP's capabilities, this work explores the complexities of facial emotion manipulation, addressing existing methodological limitations within the digital landscape. Central to our approach is the Emotion Manipulation Mapper, which incorporates additional conditional information to boost domain awareness and adapt to diverse emotional distributions. This capability ensures the creation of realistic and contextually appropriate facial expressions. Furthermore, a Dual Auxiliary Classifier enriches the model with a broad spectrum of label information, significantly heightening the precision of emotion manipulation. Additionally, the integration of a CLIP text encoder offers crucial textual insights that guide the emotion manipulation process, while a pre-trained StyleGAN2 generator ensures the creation of high-quality, visually appealing images. Compared to existing methods, our approach excels in manipulated image quality, identity preservation, and manipulation accuracy. A comprehensive set of evaluation metrics thoroughly analyzes model performance in handling facial emotion manipulation, providing valuable insights into the efficacy and potential applications of this innovative method.

Acknowledgments. This work is supported by National Natural Science Foundation of China under grant 62176062.

Disclosure of Interests. The authors declare that they have no conflict of interest.

References

1. Liao, M., Fan, X., Li, Y., Gao, M.: Noise-related face image recognition based on double dictionary transform learning. Information Sciences **630**, 98–118 (2023)
2. Drummond, J., Makdani, A., Pawling, R., Walker, S.C.: Congenital anosmia and facial emotion recognition. Physiology & Behavior **278**, 114519 (2024)
3. Azari, B., Lim, A.: EmoStyle: One-shot facial expression editing using continuous emotion parameters. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6385–6394 (2024)
4. Luvembe, A.M., Li, W., Li, S., Liu, F., Wu, X.: CAF-ODNN: Complementary attention fusion with optimized deep neural network for multimodal fake news detection. Information Processing & Management **61**(3), 103653 (2024)

5. Shen, Q., Xu, J., Mei, J., Wu, X., Dong, D.: EmoStyle: Emotion-aware semantic image manipulation with audio guidance. Applied Sciences **14**(8), 3193 (2024)
6. Liu, Y., Li, Q., Deng, Q., Sun, Z., Yang, M.-H.: Gan-based facial attribute manipulation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
7. Yauri-Lozano, E., Castillo-Cara, M., Orozco-Barbosa, L., García-Castro, R.: Generative adversarial networks for text-to-face synthesis & generation: A quantitative–qualitative analysis of natural language processing encoders for Spanish. Information Processing & Management **61**(3), 103667 (2024)
8. Mulder, M.J., Prummer, F., Terburg, D., Kenemans, J.L.: Drift–diffusion modeling reveals that masked faces are preconceived as unfriendly. Scientific Reports **13**(1), 16982 (2023)
9. Zhu, J., Mu, L.: GrainedCLIP and DiffusionGrainedCLIP: Text-guided advanced models for fine-grained attribute face image processing. IEEE Access (2023)
10. Hou, X., Shen, L., Ming, Z., Qiu, G.: Deep generative image priors for semantic face manipulation. Pattern Recognition **139**, 109477 (2023)
11. Wang, J., Zhao, K., Ma, Y., Zhang, S., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Facecomposer: A unified model for versatile facial content creation. Advances in Neural Information Processing Systems **36** (2024)
12. Sun, Z., Wen, Y.-H., Lv, T., Sun, Y., Zhang, Z., Wang, Y., Liu, Y.-J.: Continuously controllable facial expression editing in talking face videos. IEEE Transactions on Affective Computing, 1–14 (2023)
13. Liu, W., Zhang, S., Zhou, L., Luo, N., Chen, Q.: Sound to expression: Using emotional sound to guide facial expression editing. Journal of King Saud University-Computer and Information Sciences, 101998 (2024)
14. Yan, S., Soladié, C., Aucouturier, J.-J., Seguier, R.: Combining GAN with reverse correlation to construct personalized facial expressions. Plos One **18**(8), e0290612 (2023)
15. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831 (2021)
16. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2426–2435 (2022)
17. Ghandchi, A., Golbabaei, S., Borhani, K.: Effects of two different social exclusion paradigms on ambiguous facial emotion recognition. Cognition and Emotion, 1–19 (2023)
18. Chang, H., Zhang, H., Barber, J., Maschinot, A.J., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W.T., Rubinstein, M.: Muse: Text-to-image generation via masked generative transformers. In: Proceedings of the 40th International Conference on Machine Learning, pp. 4055–4075 (2023)
19. Frans, K., Soros, L., Witkowski, O.: Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. Advances in Neural Information Processing Systems **35**, 5207–5218 (2022)
20. Sowden, S., Schuster, B.A., Keating, C.T., Fraser, D.S., Cook, J.L.: The role of movement kinematics in facial emotion expression production and recognition. Emotion **21**(5), 1041 (2021)
21. Romero-Martínez, Á., Sarrate-Costa, C., Moya-Albiol, L.: A systematic review of the role of oxytocin, cortisol, and testosterone in facial emotional processing. Biology **10**(12), 1334 (2021)

22. Barzilay, N., Shalev, T.B., Giryes, R.: MISS GAN: A multi-illustrator style generative adversarial network for image to illustration translation. Pattern Recognition Letters **151**, 140–147 (2021)
23. Gal, R., Patashnik, O., Maron, H., Bermano, A.H., Chechik, G., Cohen-Or, D.: StyleGAN-NADA: CLIP-guided domain adaptation of image generators. ACM Transactions on Graphics (TOG) **41**(4), 1–13 (2022)
24. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems **34**, 17981–17993 (2021)
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
26. Wu, R., Zhang, G., Lu, S., Chen, T.: Cascade ef-gan: Progressive facial expression editing with local focuses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5021–5030 (2020)
27. Wang, Y., Zhang, Z., Hao, W., Song, C.: Multi-domain image-to-image translation via a unified circular framework. IEEE Transactions on Image Processing **30**, 670–684 (2020)
28. Strizhkova, V., Wang, Y., Anghelone, D., Yang, D., Dantcheva, A., Brémond, F.: Emotion editing in head reenactment videos using latent space manipulation. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 1–8 (2021)
29. Wang, J., Zhang, J., Lu, Z., Shan, S.: DFT-Net: Disentanglement of face deformation and texture synthesis for expression editing. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3881–3885 (2019)
30. Noor, N.A.N.M., Suaib, N.M.: Facial expression transfer using generative adversarial network: A review. In: IOP Conference Series: Materials Science and Engineering, vol. 864(1), 012077. IOP Publishing (2020)
31. Lee, C.-H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5549–5558 (2020)
32. 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. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
33. 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, pp. 8110–8119 (2020)
34. 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, pp. 4401–4410 (2019)
35. 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, pp. 2085–2094 (2021)
36. Carlier, A., Danelljan, M., Alahi, A., Timofte, R.: Deepsvg: A hierarchical generative network for vector graphics animation. Advances in Neural Information Processing Systems **33**, 16351–16361 (2020)
37. Kwon, G., Ye, J.C.: CLIPstyler: Image style transfer with a single text condition supplementary materials (n.d.)
38. Tulyakov, S., Liu, M.-Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1526–1535 (2018)

39. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: One-shot anatomically consistent facial animation. International Journal of Computer Vision **128**, 698–713 (2020)
40. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.-P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66 (2018)
41. 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, pp. 9243–9252 (2020)
42. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(4), 2004–2018 (2020)
43. Xia, W., Yang, Y., Xue, J.-H., Wu, B.: Tedigan: Text-guided diverse face image generation and manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2256–2265 (2021)
44. Kocasari, 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, pp. 895–904 (2022)
45. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., Van Knippenberg, A.D.: Presentation and validation of the Radboud Faces Database. Cognition and Emotion **24**(8), 1377–1388. Taylor & Francis (2010)
46. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) **40**(4), 1–14. ACM New York, NY, USA (2021)
47. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
48. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: International Conference on Machine Learning, pp. 2642–2651. PMLR (2017)
49. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
50. Li, B., Luo, S., Qin, X., Pan, L.: Improving gan with inverse cumulative distribution function for tabular data synthesis. Neurocomputing **456**, 373–383. Elsevier (2021)
51. 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 **10**(1), 18–31 (2017)
52. Rodríguez-Fuertes, A., Alard-Josemaría, J., Sandubete, J.E.: Measuring the candidates' emotions in political debates based on facial expression recognition techniques. Frontiers in Psychology **13**, 785453 (2022)
53. Zhu, D., Fu, Y., Zhao, X., Wang, X., Yi, H.: Facial emotion recognition using a novel fusion of convolutional neural network and local binary pattern in crime investigation. Computational Intelligence and Neuroscience **2022**. Hindawi Limited (2022)
54. Sivaiah, B., Gopalan, N.P., Mala, C., Lavanya, S.: FL-CapsNet: Facial localization augmented capsule network for human emotion recognition. Signal, Image and Video Processing **17**(4), 1705–1713. Springer (2023)

55. Valente, A., Lopes, D.S., Nunes, N., Esteves, A.: Empathic AuRea: Exploring the effects of an augmented reality cue for emotional sharing across three face-to-face tasks. In: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 158–166 (2022)
56. Sun, Z., Wen, Y.-H., Lv, T., Sun, Y., Zhang, Z., Wang, Y., Liu, Y.-J.: Continuously controllable facial expression editing in talking face videos. IEEE Transactions on Affective Computing (2023)
57. Cheng, K., Cun, X., Zhang, Y., Xia, M., Yin, F., Zhu, M., Wang, X., Wang, J., Wang, N.: Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In: SIGGRAPH Asia 2022 Conference Papers, pp. 1–9 (2022)
58. Sun, J., Wang, X., Zhang, Y., Li, X., Zhang, Q., Liu, Y., Wang, J.: Fenerf: Face editing in neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7672–7682 (2022)
59. Hou, X., Zhang, X., Liang, H., Shen, L., Lai, Z., Wan, J.: Guidedstyle: Attribute knowledge guided style manipulation for semantic face editing. Neural Networks **145**, 209–220. Elsevier (2022)
60. Abdal, R., Zhu, P., Femiani, J., Mitra, N., Wonka, P.: Clip2stylegan: Unsupervised extraction of stylegan edit directions. In: ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–9 (2022)
61. Zhao, C., Cai, W.-L., Yuan, Z.: Spectral normalization and dual contrastive regularization for image-to-image translation. The Visual Computer, 1–12. Springer (2024)
62. Zahara, L., Musa, P., Wibowo, E., Karim, I., Musa, S.B.: The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi. In: 2020 Fifth International Conference on Informatics and Computing (ICIC), pp. 1–9 (2020)