Learning Adaptive Spatial Coherent Correlations for Speech-Preserving Facial Expression Manipulation

Tianshui Chen
tianshuichen@gmail.com

Jianman Lin
linjianmancjx@gmail.com

Zhijing Yang*
yzhj@gdut.edu.cn

Chunmei Qing
qchm@scut.edu.cn

Liang Lin
linliang@ieee.org

Abstract

Speech-preserving facial expression manipulation (SPFEM) aims to modify facial emotions while meticulously maintaining the mouth animation associated with spoken content. Current works depend on inaccessible paired training samples for the person, where two aligned frames exhibit the same speech content yet differ in emotional expression, limiting the SPFEM applications in real-world scenarios. In this work, we discover that speakers who convey the same content with different emotions exhibit highly correlated local facial animations, providing valuable supervision for SPFEM. To capitalize on this insight, we propose a novel adaptive spatial coherent correlation learning (ASCCL) algorithm, which models the aforementioned correlation as an explicit metric and integrates the metric to supervise manipulating facial expression and meanwhile better preserving the facial animation of spoken contents. To this end, it first learns a spatial coherent correlation metric, ensuring the visual disparities of adjacent local regions of the image belonging to one emotion are similar to those of the corresponding counterpart of the image belonging to another emotion. Recognizing that visual disparities are not uniform across all regions, we have also crafted a disparity-aware adaptive strategy that prioritizes regions that present greater challenges. During SPFEM model training, we construct the adaptive spatial coherent correlation metric between corresponding local regions of the input and output images as addition loss to supervise the generation process. We conduct extensive experiments on variant datasets, and the results demonstrate the effectiveness of the proposed ASCCL algorithm. Code is publicly available at https://github.com/jianmanlincjx/ASCCL

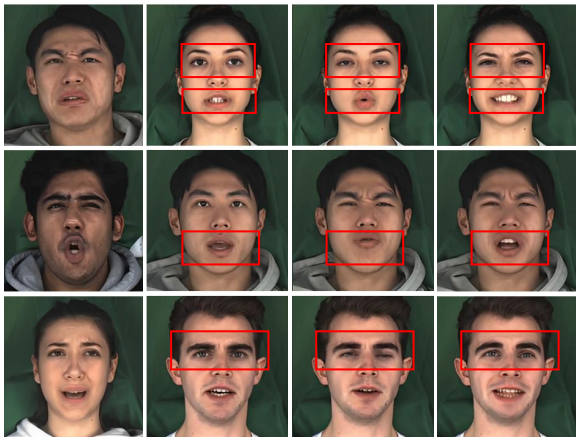

Figure 1. Several examples are generated by the current advanced NED with and without the proposed ASCCL algorithm. Incorporating the ASCCL can better manipulate the expressions and meanwhile preserve mouth shapes.

1. Introduction

Speech-preserving facial expression manipulation (SPFEM), which aims to manipulate facial emotions while preserving the mouth animations in static images or dynamic videos, can enhance human expressiveness and thus benefit variant applications including virtual avatars and film & television production. For example, it requires lots of effects and repeated remakes to capture an expected actor's emotions in a movie & shooting. In contrast, a robust SPFEM system can easily modify the facial emotions to achieve comparable performance in the post-production stage and thus is urgently expected.

Current SPFEM literature either predominantly previous face reenactment algorithms [10, 28] or harnesses decou-

*Zhijing Yang is the corresponding author. Tianshui Chen, Jianman Lin, and Zhijing Yang are with Guangdong University of Technology. Chunmei Qing is with South China University of Technology. Liang Lin is with Sun Yat-Sen University. This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant No. 62206060, in Part by Natural Science Foundation of Guangdong Province (2022A1515011555, 2023A1515012568, 2023A1515012561), Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004), and in part by Guangzhou Basic and Applied Basic Research Foundation under Grant No. SL2022A04J01626.

pled semantic representations equipped by cyclic consistency [23, 36]. The former category of works [10, 28] typically manipulates facial expressions through the exchange of latent codes [17] or facial action units [13], and employs the reference images as surrogate labels to construct frame-by-frame construction supervision. However, these surrogate images are not perfect representations of the desired outcomes and the reliance on them may lead to generating sub-optimal results. The latter approach [23] posits a one-to-one correspondence between images exhibiting varied emotional expressions and employs cyclic consistency for paired supervision. Despite achieving better performance, the global cyclical consistency constraint makes it difficult to capture alterations in fine-grained facial information under different emotions. Consequently, these methods either fail to accurately translate the intended emotions (as illustrated in the first and third examples in Figure 1) or do not maintain the original mouth animation tied to the spoken content (as evidenced in the first and second examples in Figure 1).

To establish accurate information for additional supervisory guidance, we investigate a plausible and reasonable assumption: a single speaker articulating identical content across varied emotional states exhibits a high degree of correlation in local facial animations. Within the realm of the SPFEM task, we posit that there exists a strong correlation between adjacent local regions of the input images and their corresponding counterparts in the output images. For a more detailed analysis, we delve into the examination of similarities and correlation coefficients between corresponding local regions of the input and output images, and those between non-corresponding local regions of the input and output images. As depicted in Figure 2, there is a notable presence of high similarities and correlation coefficients for corresponding local regions. Conversely, the non-corresponding local regions demonstrate values approaching zero. These observations provide robust empirical support for the previously stated assumption and underscore the merit of modeling these correlations to facilitate SPFEM performance.

In this work, we design a novel adaptive spatial coherent correlation learning (ASCCL) algorithm, which discerns the correlations between adjacent local regions within images that display varying emotional states and incorporates these correlations as additional guidance to supervise the manipulation of expressions in a manner that is attuned to difficulty. Formally, we leverage visual disparities as a means to characterize the interactions between adjacent local regions, given that local motion disparities are critical to the realism of facial animations. Then, we formulate a spatial coherent correlation metric ensuring the visual disparities of adjacent local regions of the image belonging to one emotion are similar to those of the corresponding counterpart of the image belonging to another emotion. Recognizing the variable complexity across different facial regions, we introduce a disparity-aware adaptive strategy, which preferentially weights more challenging regions with higher values while proportionally reducing the weight for less complex areas. During SPFEM model training, we establish a network of dense correlations between local regions of the input and output images, employing this adaptive spatially coherent correlation metric as an auxiliary supervision. ASCCL is trained using a set of paired data and can be used to construct supervision for any other persons to facilitate generating high-quality results in a plug-and-play manner.

The contributions can be summarized into three folds. Firstly, we introduce an adaptive spatial coherent correlation learning (ASCCL) algorithm, which learns the consistent correlations between input and generated images in terms of the visual disparities of adjacent local regions. It can be seamlessly integrated into current SPFEM methods to improve their performance in a plug-and-play manner. Secondly, we introduce a difficult-aware adaptive strategy that weights more challenging regions with higher constraints while proportionally reducing the weight for less complex areas. Lastly, we conduct extensive experiments that integrate the ASCCL algorithms into current advanced methods, demonstrating the effectiveness of the proposed ASCCL algorithm.

2. Related Works

Video-based face manipulation. To modify talking/moving faces, video-based face manipulation algorithms frequently employ conditionally Generative Adversarial Networks (GANs) [14, 19, 29, 31, 33] or 3DMM [2, 9, 11, 12, 27]. For example, GANimation [25] leverages adversarial learning conditioning on action unit [13] annotations to describe facial movements in a continuous manifold. [29] apply StyleGAN [18] to learn low-frequency information to accomplish temporal coherency. DSM [26] learns disentangled representation and controls facial expressions via semantic representation. SPFEM is more difficult than simple face manipulation since it requires not only modifying facial expressions but also retaining the facial motion of original speech contents.

Face reenactment. Facial reenactment, in which a specific actor imitates speech and expressions from a reference video [4, 8, 10, 28, 34] and other [35]. For example, ICface [28] controls the pose and expression with interpretable control signals such as head pose angles and action units. To create temporally consistent videos, Head2Head++ [10] employs a sequential generator and a customized dynamics discriminator. StyleHEAT [34] extends the latent code of StyleGAN [17] to aid in motion and expression generation and animation. SPFEM is similar to this task in expression

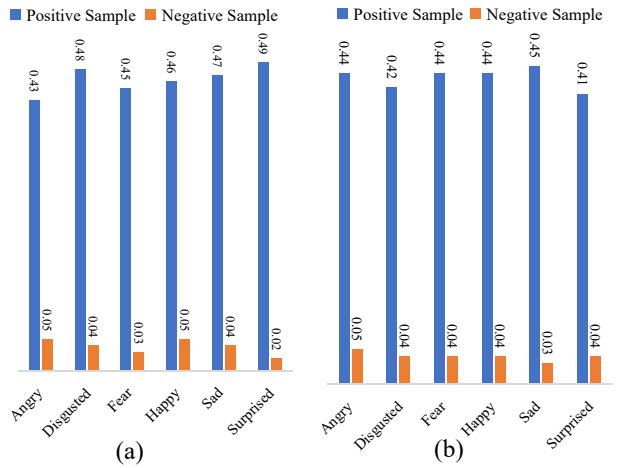

Figure 2. (a) Average similarities and (b) Average Pearson correlation coefficients of the positive and negative samples.

manipulation, but it also requires preserving mouth movement of original speech contents, making SPFEM a more difficult task.

Speech-preserving facial expression manipulation. SPFEM seeks to alter the given source video to the desired emotion while keeping the voice content's facial animation. Previous research has adapted facial reenactment algorithms such as ICface [28] to this job. However, facial reenactment can keep both the expression and the mouth form, but not the speech substance. To accomplish this challenge, [23] propose combining the 3DMM parameters of source identification and target emotion. Despite making significant progress, these works lack paired supervision, resulting in sub-optimal performance in emotion manipulation and speech preservation. Different from the above-mentioned works, we present in-depth analyses about the spatial coherent correlations across images belonging to different emotions and find there exist strong correlations between adjacent local regions of the input image of one emotion and their corresponding counterparts in the output image of another emotion. We propose to learn adaptive spatial coherent correlations to construct additional supervision for SPFEM.

3. Motivation

As discussed above, we make a plausible and reasonable assumption that a single speaker articulating identical content across varied emotional states exhibits a high degree of correlation in local facial animations. To present a more in-depth and direct analysis, we further conduct statistical experiments to validate this assumption.

Formally, given two images of the same speaker expressing the same content with two different emotions, denoted as x and y, we use the output of the last convolution layer of pre-trained network [7] to compute their feature x^f and y^f. Here, we use visual disparity to denote the relationship of regions at locations i and j, and the similarity between visual disparities of corresponding and non-corresponding adjacent local regions, formulated as:

$$\begin{aligned} \mathrm{s}^p_{i\to j} &= \varphi(x^f_{i\to j}, y^f_{i\to j}) \\ \mathrm{s}^n_{i\to k} &= \varphi(x^f_{i\to j}, y^f_{i\to k, k\neq j}) \end{aligned} \tag{1}$$

$x^f_{i\to j}$ and $y^f_{i\to j}$ denote visual disparities of regions i and j of the images x and y. φ represents the cosine similarity function. For each emotion pair, we retrieve thousands of corresponding and non-corresponding adjacent local regions and compute their average similarities. As shown in Figure 2 (a), we present the average similarities of corresponding and non-corresponding adjacent local regions according to the emotion pairs of neural to other six emotions. We find the average similarities of the non-corresponding adjacent local regions approach 0, and that of the corresponding adjacent local regions range from 0.43 to 0.49. Moreover, we further compute Pearson correlation coefficients [6] between visual disparities of visual disparities of corresponding and non-corresponding adjacent local regions. Similarly, we use thousands of corresponding and non-corresponding adjacent local regions and compute their average Pearson correlation coefficients. As presented in Figure 2 (b), a similar phenomenon is observed. These results suggest there exist strong correlations between the corresponding adjacent local regions between two images expressing the same content while differing in emotions. Thus, it is worth investigating to learn the correlations and integrate the correlations to facilitate the SPFEM performance.

4. Method

It first learns the spatially coherent correlation metric, ensuring the alignment between the visual disparity of adjacent local regions of the source input image and that of the corresponding counterpart of the output generated image. Then, it designs a disparity-aware adaptive strategy to weigh more challenging regions with higher values while reducing the weight for less complex regions. The ASCCL algorithm can be used to supervise generating visual contents, including intermediate results (e.g., 3DMM [1]) and final rendered images in a plug-and-play manner. An overall pipeline for incorporating ASCCL into the two-stage NED [23] method is illustrated in Figure 3.

4.1. Spatial Coherent Correlation Metric Learning

As suggested in Section 3, there inherently exist strong correlations between adjacent regions of two images that one speaker expresses the same content while differing in different emotions. Indeed, we can further strengthen the correlations by learning the spatial coherent correlation (SCC)

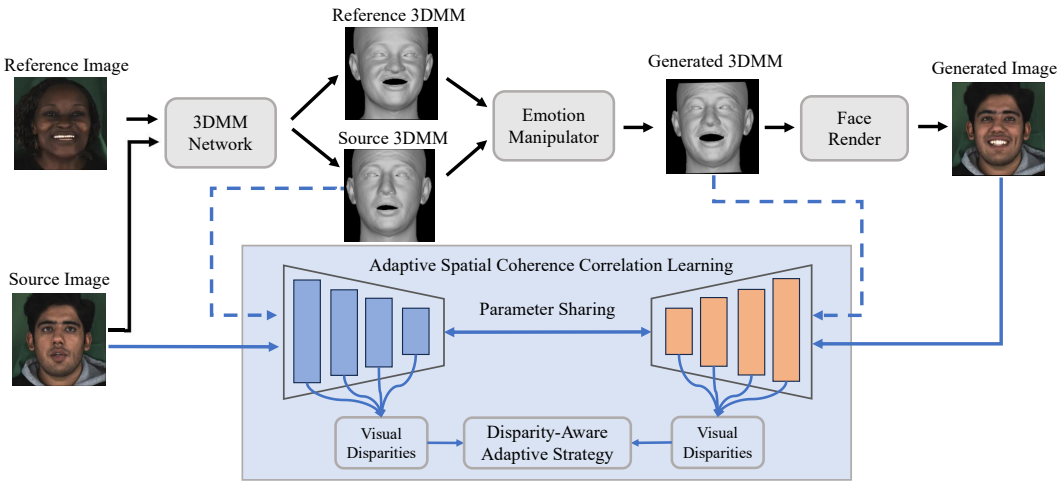

Figure 3. An overall pipeline of incorporating the proposed ASCCL algorithm to the current advanced NED [23] method to supervise generating the intermediate 3DMM meshes and final rendered images. It computes visual disparities of corresponding and non-corresponding local regions between the source and generated images, followed by the disparity-aware adaptive strategy to obtain the final loss to supervise final image generation. An identical process is performed on the source and generated 3DMM meshes to supervise the intermediate 3DMM mesh generation.

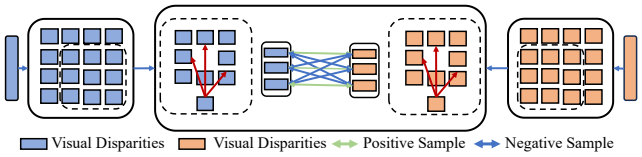

Figure 4. An illustration of spatial coherent correlation metric learning. It retrieves the corresponding adjacent local regions of the input and output images as positive samples and the non-corresponding counterparts as negative samples. The process is performed in the feature maps to construct dense positive and negative samples to train the metric.

metric using paired data, which can assign a higher value to the correlation between corresponding adjacent regions and assign a smaller value to the correlation between non-corresponding counterparts. In this way, it can provide better-synchronized signals to supervise the generation process. Here, we introduce the SCC metric learning in detail.

Formally, given two images of the same speaker expressing the same content with two different emotions, we need to extract corresponding regions i and j of both two images. However, pose bias typically exists between these two images, potentially compromising the calculation accuracy. To address this issue, we first introduce a pose alignment module that extracts landmark coordinates using a pre-trained model in the OpenCV library for both two images and utilizes these coordinates to compute an affine transformation matrix. This matrix is subsequently applied to align the pose information between the two images and obtain two images x and y that ensure geometric correspondence.

Figure 4 exhibits the detailed computation process of the SCC metric. Owing to the locality and translation invariance characteristics of convolutional neural networks [7, 15], the mapping of pixels in the feature space back to the original image is approximately equivalent to a local region, and as the depth of the network increases, the receptive field also expands. Leveraging these characteristics, we employ the output of multiple convolutional layers to obtain the multi-scale features. Formally, we can extract the feature of layer l as:

$$\begin{aligned} x^{f,l} &= \phi^l(x) \\ y^{f,l} &= \phi^l(y) \end{aligned} \tag{2}$$

Here, $\phi^l(\cdot)$ is feature extractor for layer l. We adopt the Arcface network [7], and use the output of four convolutional layers with dimension decline and thus $l \in [1, 2, 3, 4]$. Given two regions i and j, we compute the visual disparity using a mapping of the feature difference, denoted as:

$$\begin{aligned} x^{f,l}_{i\to j} &= f(x^{f,l}_i - x^{f,l}_j) \\ y^{f,l}_{i\to j} &= f(y^{f,l}_i - y^{f,l}_j) \end{aligned} \tag{3}$$

where $f(\cdot)$ denotes a mapping function and it is implemented using two stacked fully-connected layers cooperated with the rectified linear unit non-linear function. Inspired by the recent progress in previous works [3, 16, 22, 32], we define an identical contrastive loss, formulated as:

$$\ell^{l,i\to j}_{xy} = \frac{\exp(\frac{x^{f,l}_{i\to j}\cdot y^{f,l}_{i\to j}}{\tau})}{\exp(\frac{x^{f,l}_{i\to j}\cdot y^{f,l}_{i\to j}}{\tau}) + \exp(\sum_{k=1,k!=j}^{m} \frac{x^{f,l}_{i\to j}\cdot y^{f,l}_{i\to k}}{\tau})} \tag{4}$$

where τ stands for a temperature hyper-parameter set to 0.07 by default. The final loss can be defined as the summation of all image pairs, corresponding region pairs, and four

layers, formulated as:

$$\mathcal{L}_{sccl} = \sum_{l=1} \sum_{i,j} \sum_{x,y} \ell_{xy}^{l,i\to j} \tag{5}$$

By minimizing losses through backpropagation, we can align the visual disparities of y to x with different emotions, and thus provide visual consistency correlations to supervise SPFEM model training.

4.2. Disparity-Aware Adaptive Strategy

Once the SCCL metric is learned, we can use to $\mathcal{L}_{sccl}$ loss between the input source (source 3DMM) and output generated (generated 3DMM) images as additional supervision. Here, we observe a notable phenomenon that the complexity varies across different facial regions. For example, it is more complex and challenging for the mouth regions as it change dramatically when the speaker is talking. In contrast, when the regions are positioned at the cheek area, the visual disparity is reduced, promoting a faster convergence. This underscores the region-specific sensitivity of visual disparity. Consequently, by enabling the SCC metric to independently discern and dynamically modulate learning strategies for specific regions, we can enhance the extraction of spatially coherent correlation information in an adaptive manner.

Inspired by the idea of [20], DAAS is engineered to enable the SCC metric to discern this variability and tailor its learning strategies based on the characteristics of the regions, prioritizing those with higher learning complexity. Specifically, when the visual disparity $x_{i\to j}^{f,l}$ is larger, this region is deemed a challenging region, to which we assign a larger weight value. Conversely, when $x_{i\to j}^{f,l}$ is small, it is considered a simple region that can achieve rapid convergence, we reduce the weight for those simple regions. To this end, we propose to assign different weights according to the visual disparity, formulated as:

$$w_{ij}^{l} = \lambda \cdot \text{sigmoid}(x_{i\to j}^{f,l})^{r} \tag{6}$$

where w_{ij}^{l} represents the weight value of adjacent regions $i \to j$ for images x and y at the feature map layer l. λ and r are hyper-parameters, both of which are set to 2 to ensure a reasonable weight. The final loss function can be defined as:

$$\mathcal{L}_{asccl} = \sum_{l=1} \sum_{i,j} \sum_{x,y} w_{ij}^{l} \cdot \ell_{xy}^{l,i\to j} \tag{7}$$

Current SPFEM algorithms can be divided into two types. The first involves a two-stage generation: initially creating 3DMM parameters and then using them to render final images [23]. The second method directly produces rendered images [28]. Visual information here refers to either the 3DMM parameters or the final images. $\mathcal{L}_{asccl}$ can be used for both these two types of algorithms as shown in Figure 3. *Due to page limit, we present more implementation details, including network architectures, integration to current NED and ICface method, and training details.*

5. Experiments

5.1. Dataset

We performed experiments on the MEAD dataset [30], which contains 60 speakers, and each speaker records 30 videos in each emotional state (i.e., neutral, happy, angry, surprised, fear, sad, and disgusted). Here, we selected videos of 36 speakers that have 7,560 videos to train the ASCCL algorithm. To evaluate the SPFEM model's performance, we select 6 non-overlapped speakers (M003, M009, W029, M012, M030, and W015) that have 1,260 videos. We randomly selected 90% as the training set and the rest 10% as the test set similar to previous works [23]. We additionally employ the ASCCL algorithm on the well-established RAVDESS dataset [21] without the need for retraining. Specifically, we focus on 6 speakers (actors 1-6) encompassing 168 videos. Similarly, 90% of the videos are randomly chosen for the training set, while the remaining 10% constitute the test set.

5.2. Evaluation Protocol

In this work, we use these metrics for evaluation: 1) Frechet Arcface Distance (FAD) gauges video realism by comparing feature vectors of generated and real videos using advanced face recognition technology [7]. Lower FAD values indicate better realism. 2) Cosine Similarity (CSIM) assesses emotional similarity between generated and target emotional videos using a state-of-the-art expression recognition network, with higher CSIM values denoting greater similarity. 3) Lip Sync Error Distance (LSE-D) [24] evaluates lip-audio accuracy using a pre-trained model [5] to measure the disparities between lip and audio representations, with smaller LSE-D indicates a higher lip-audio accuracy. We present the results of two settings: inter-identification, where emotion reference and source video share the same speaker, and cross-identification, involving different speakers.

5.3. Comparison with Baseline Methods

5.3.1 Quantitative Comparisons

We first present the performance comparisons on MEAD in Table 1. When ASCCL is integrated into NED, the resulting image sequences are greatly improved in FAD, LSE-D, and CSIM. In the Cross-ID setting, compared with NED itself, FAD, LSE-D, and CSIM have all improved to a certain extent, with FAD reduced from 4.448 to 4.264, LSE-D from 9.906 to 9.238, and CSIM from 0.773 to 0.791. Among

Settings	Emotions	ICface			Ours (ICface)			NED			Ours (NED)		
		FAD↓	LSE-D↓	CSIM↑	FAD↓	LSE-D↓	CSIM↑	FAD↓	LSE-D↓	CSIM↑	FAD↓	LSE-D↓	CSIM↑
Inter-ID	Neutral	7.114	9.760	0.779	7.158	9.382	0.781	0.906	9.264	0.883	0.891	9.113	0.911
	Angry	6.420	10.483	0.741	6.195	9.766	0.803	2.177	9.579	0.802	1.195	9.550	0.896
	Disgusted	7.383	10.433	0.805	6.265	9.266	0.809	3.838	9.128	0.772	1.679	9.416	0.854
	Fear	6.567	9.855	0.754	6.698	9.481	0.800	1.659	10.172	0.848	1.330	10.039	0.900
	Happy	6.213	10.180	0.775	6.191	9.556	0.837	1.939	9.137	0.839	1.326	8.975	0.930
	Sad	7.301	10.017	0.755	6.727	9.293	0.779	2.538	9.074	0.812	1.162	8.906	0.906
	Surprised	6.567	9.851	0.817	7.470	9.271	0.795	1.700	9.821	0.864	1.056	9.382	0.902
	Avg.	6.795	10.083	0.775	6.672	9.431	0.801	2.108	9.454	0.831	1.234	9.340	0.900
Cross-ID	Neutral	10.560	11.226	0.705	9.745	10.604	0.671	2.022	9.812	0.841	1.865	9.134	0.852
	Angry	9.470	11.073	0.648	9.271	10.456	0.693	4.851	9.904	0.717	4.853	9.239	0.748
	Disgusted	9.230	11.184	0.637	9.323	10.338	0.764	5.094	10.121	0.791	4.840	9.347	0.814
	Fear	9.122	11.204	0.727	9.273	10.221	0.729	4.983	9.741	0.750	4.820	9.239	0.761
	Happy	8.493	11.322	0.717	9.505	10.421	0.793	3.919	9.936	0.842	3.383	9.178	0.866
	Sad	10.364	11.526	0.664	9.710	10.314	0.664	5.665	10.179	0.691	5.475	9.427	0.720
	Surprised	9.541	11.133	0.721	9.534	10.273	0.766	4.600	9.646	0.780	4.615	9.105	0.779
	Avg.	9.540	11.238	0.688	9.480	10.375	0.726	4.448	9.906	0.773	4.264	9.238	0.791

Table 1. Comparision results of FAD, CSIM, and LSE-D of NED, ICFACE with and without our ASCCL on the inter-IDentification and cross-IDentification settings on the MEAD dataset.

them, the LSE-D has seen significant improvements, and these notable advancements in the LSE-D metric can be attributed to ASCCL's pivotal role in supervising the training of the SPFEM model. This supervision accentuates the intrinsic visual correlations between the SPFEM model's input and output, ensuring the visual disparity of the input and output align closely thereby enhancing consistency in mouth shape modifications related to emotion. Additionally, the betterment in FAD and CSIM further attests that ASCCL not only maintains but can also elevate the model's training quality, steering it towards superior optimization. In the inter-ID setting, these three indicators are also improved to a certain extent, indicating that the proposed ASCCL algorithm has strong generalization and can adapt to the emotional migration of inter-ID and cross-ID. When incorporating the ASCCL into the single-stage method ICface can also obtain significant performance improvement. In the Inter-ID setting, compared with ICface itself, FAD, LSE-D, and CSIM have all improved to a certain extent, with FAD reduced from 6.795 to 6.672, LSE-D from 10.083 to 9.431, and CSIM from 0.775 to 0.801. Cross-ID is a more general and practical setting, and incorporating the ASCCL also achieves evident improvement on FAD, LSE-D, and CSIM of different expression manipulation as well as the average FAD, LSE-D, and CSIM as shown in Table 1

To demonstrate the generalization ability of the trained ASCCL, we also present the performance comparisons on the RAVDESS [21] dataset without retraining ASCCL. As shown in Table 2, incorporating ASCCL can also obtain obvious improvement for different expression manipulation in both settings. when using the ICface baseline in the inter-ID setting, it decreases the average FAD, LSE-D by 1.142, 0.436, and increases the average CSIM by 0.01. In the Cross-ID setting, it decreases the average FAD, LSE-D by 0.521, 1.226, and the average CSIM remains the same. Similar performance improvement is obtained when using the NED baseline. The experiment corroborates ASCCL's robustness across different methods and data domains.

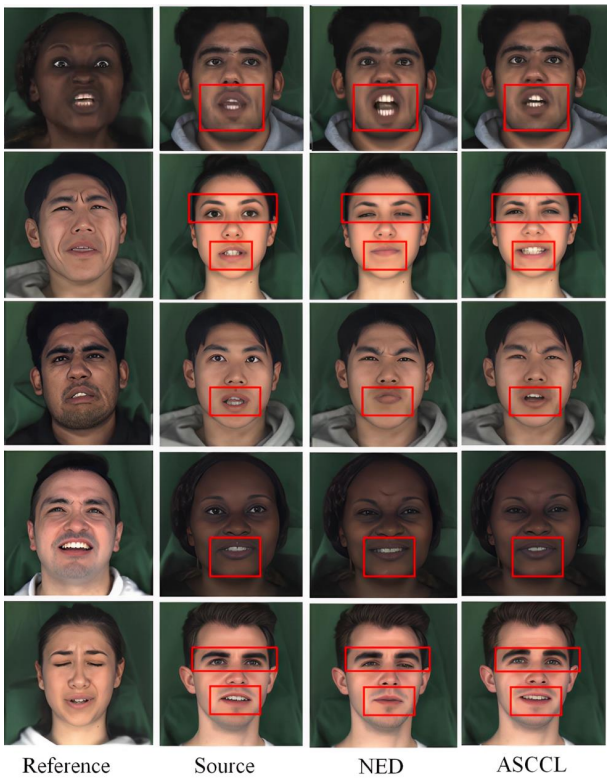

Figure 5. Qualitative comparisons of NED with and without the proposed ASCCL algorithm. The samples are selected from the MEAD dataset.

5.3.2 Qualitative Comparisons

In this section, we exhibit some visualization results of the baseline NED methodology, both with and without the ASCCL algorithm, as illustrated in Figure 5. Analogous to the quantitative metrics, we also dissect the qualitative comparisons from three dimensions. 1) Realism: The eye region tends to be more closed as a consequence of the NED [23] application, as exemplified in the third column of the second and fifth rows. Furthermore, the mouth region is distorted due to the imprecise prediction of the mouth shape as

Settings	Emotions	ICface			Ours (ICface)			NED			Ours (NED)		
		FAD↓	LSE-D↓	CSIM↑	FAD↓	LSE-D↓	CSIM↑	FAD↓	LSE-D↓	CSIM↑	FAD↓	LSE-D↓	CSIM↑
Inter-ID	Neutral	9.816	8.209	0.749	8.302	7.239	0.762	2.041	7.376	0.847	2.810	7.567	0.853
	Angry	7.047	9.504	0.703	6.182	9.453	0.711	3.288	7.757	0.805	3.722	7.601	0.799
	Disgusted	8.689	8.295	0.775	7.066	8.649	0.783	4.144	7.822	0.786	3.194	8.100	0.836
	Fear	8.413	8.523	0.722	7.406	8.625	0.741	2.635	7.452	0.842	2.588	7.821	0.831
	Happy	8.413	8.902	0.797	7.403	7.924	0.797	3.714	7.742	0.793	3.025	6.678	0.826
	Sad	8.086	8.346	0.766	7.337	7.018	0.781	2.595	7.560	0.855	2.525	7.086	0.842
	Surprised	8.636	7.578	0.772	7.411	7.402	0.781	2.980	7.226	0.848	3.410	7.299	0.824
	Avg.	8.443	8.480	0.755	7.301	8.044	0.765	3.057	7.562	0.825	3.039	7.450	0.830
Cross-ID	Neutral	10.478	10.736	0.677	10.343	9.302	0.682	3.558	7.856	0.820	3.160	7.458	0.813
	Angry	8.704	12.415	0.646	8.242	10.522	0.706	5.546	8.085	0.766	4.851	8.513	0.739
	Disgusted	9.260	11.860	0.717	8.948	10.234	0.652	7.388	8.107	0.741	7.443	7.848	0.689
	Fear	9.106	11.279	0.649	9.062	10.382	0.753	5.008	8.151	0.749	4.160	7.842	0.799
	Happy	9.061	11.150	0.738	9.063	9.706	0.676	5.648	8.073	0.804	4.910	7.951	0.796
	Sad	9.639	11.305	0.666	9.043	9.916	0.634	5.588	8.006	0.726	4.847	7.545	0.741
	Surprised	9.718	12.028	0.644	8.903	10.313	0.677	5.145	7.962	0.713	4.284	7.581	0.748
	Avg.	9.424	11.539	0.677	8.903	10.313	0.677	5.412	8.034	0.760	4.808	7.820	0.761

Table 2. Comparision results of FAD, CSIM, and LSE-D of NED, ICFACE with and without our ASCCL on the inter-IDentification and cross-IDentification settings on the RAVDESS dataset.

shown in the third column for the first rows. The ASCCL ameliorates this NED shortcoming by aligning the visual consistency between inputs and outputs, as demonstrated in the fourth column. 2) Emotional Similarity: Existing works fall short in manipulating facial expressions to communicate the emotions of the reference subject; they generally favor direct replication of facial components into the source, as depicted in the third column in the second, and fifth rows. The ASCCL, by maximizing the alignment of visual disparities between input and output, directs the model to prioritize the extraction of emotional information rather than merely duplicating facial components from the reference into the source, as demonstrated in the fourth column. By incorporating the ASCCL, the NED methodology is more proficient in preserving the source's contours while accomplishing emotional transference. 3) Lip-Audio Preserving Accuracy: The proposed DAAS is designed to focus on regions around with mouth area, which exhibits considerable changes in the input and output of the SFPEM model. Thus, the ASCCL can achieve consistency of the mouth shape by constraining the maximum consistency of the visual disparities for this particular region in the positive samples, as depicted in Figure 5. The results presented in the fourth column demonstrate a superior ability to retain the source's mouth shape compared to those in the third column. *We will represent more visualization results of NED and ICface with and without ASCCL algorithm on the MEAD and RAVDESS dataset in the Supplementary materials. We also present some video comparisons for more direct comparison in the Supplementary materials*

5.3.3 User study

We conducted web-based user studies to compare the performance of NED with and without the ASCCL algorithm. The study comprises three segments corresponding to the previously mentioned metrics: realism, emotion similarity with the reference emotion, and mouth shape similarity with the source video, covering seven basic emotions.

Emotion	Realism		Emotion similarity		Mouth shape similarity	
	NED	ASCCL	NED	ASCCL	NED	ASCCL
Neutral	28%	72%	24%	76%	31%	69%
Angry	28%	72%	36%	64%	30%	70%
Disgusted	35%	65%	34%	66%	26%	74%
Fear	28%	72%	35%	65%	29%	71%
Happy	28%	72%	33%	67%	28%	72%
Sad	35%	65%	28%	72%	25%	75%
Surprised	30%	70%	27%	73%	24%	76%
Avg.	30%	70%	31%	69%	27%	73%

Table 3. Realism, emotion similarity, and mouth shape similarity ratings of the user study on NED and our ASCCL.

For each emotion, we carefully selected 10 videos for both inter-identification and cross-identification settings, totaling 70 videos. Involving 25 participants, each participant was tasked with assessing the three aspects of each video. As detailed in Table 3, the inclusion of the ASCCL algorithm consistently outshines the baseline NED method across all seven emotions in all three metrics. On average, the integration of the ASCCL algorithm demonstrates significant improvement, achieving a 40% higher rating in realism, a 38% higher rating in emotion similarity, and an impressive 46% higher rating in mouth shape similarity compared to the NED baseline. *Supplementary materials include user studies utilizing the ICface baseline on MEAD and employing both NED and ICface baselines on RAVDESS.*

5.4. Ablation Study

In the ablation experiment section, we first investigate the effect of pre-training ASCCL with paired data on the final result. Following that, the pre-trained ASCCL is integrated into the NED, with an analysis of supervision on different results(3DMM or render image) and an examination of the impact of DAAS.

5.4.1 Analyses of ASCCL metric learning

Our statistical analysis demonstrates an inherent spatially coherent correlation between the input and output of the SPFEM model. Even untrained, the ASCCL can approximate this correlation, as depicted in Figure 2. By pre-

training ASCCL with paired data, we enhance the ASCCL's ability to discern inherent visual consistency. As shown in Figure 6, there is a notable improvement in the similarity and correlation coefficients for positive samples. Furthermore, we employ both trained and untrained ASCCL to guide the SPFEM model's training, as depicted in Table 4. Irrespective of whether the ASCCL has been pre-trained, it can serve as a guide when integrated into the SFPEM model. The pre-trained ASCCL demonstrated superior performance to the untrained ASCCL on measures including FAD, LSE-D, and CSIM. This indicates that by utilizing paired data to train the ASCCL, we can more effectively capture the spatial coherent correlations between the SFPEM's input and output, thereby enhancing the model's guidance.

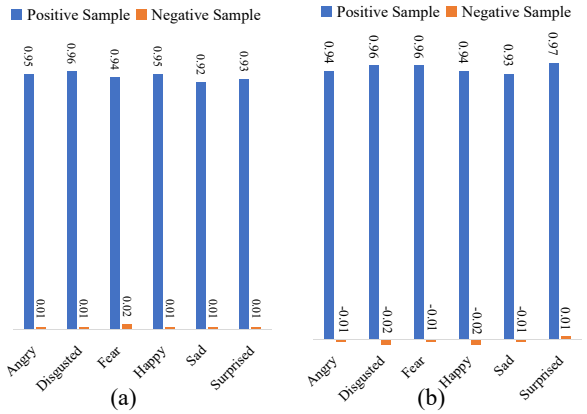

Figure 6. (a) Average similarities and (b) Average Pearson correlation coefficients of the positive and negative samples after training the SCCL metric

Settings	Methods	FAD↓	LSE-D↓	CSIM↑
Inter-ID	NED	2.108	9.454	0.831
	Ours untrained	1.509	9.434	0.861
	Ours trained	1.234	9.340	0.900
Cross-ID	NED	4.448	9.906	0.773
	Ours untrained	4.306	9.458	0.784
	Ours trained	4.264	9.238	0.791

Table 4. The performance of ASSCL with and without training.

5.4.2 Analysis of supervision on different outputs

ASCCL can supervise either the intermediate 3DMM parameters or the final rendered images. In this section, we present two additional baselines that solely utilize the objective on the 3DMM parameters (Ours 3DMM) or final rendered images (Ours Image). As depicted in Table 5, incorporating the objective solely to either 3DMM parameters or images results in significant improvements across various metrics. Moreover, introducing the objective to both the 3DMM parameters and rendered images yields even greater improvement.

Settings	Methods	FAD↓	LSE-D↓	CSIM↑
Inter-ID	NED	2.108	9.454	0.831
	Ours 3DMM	1.332	9.388	0.887
	Ours image	1.313	9.394	0.889
	Ours	1.234	9.340	0.900
Cross-ID	NED	4.448	9.906	0.773
	Ours 3DMM	4.309	9.311	0.787
	Ours image	4.276	9.321	0.785
	Ours	4.264	9.238	0.791

Table 5. Comparision results of average FAD, CSIM, and LSE-D of NED, with supervision on different outputs.

5.4.3 Analysis of Disparity-Aware Adaptive Strategy

Settings	Methods	FAD↓	LSE-D↓	CSIM↑
Inter-ID	NED	2.108	9.454	0.831
	Ours w/o DAAS	1.342	9.402	0.885
	Ours DAAS	1.234	9.340	0.900
Cross-ID	NED	4.448	9.906	0.773
	Ours w/o DAAS	4.318	9.330	0.786
	Ours DAAS	4.264	9.238	0.791

Table 6. Comparision results of average FAD, CSIM, and LSE-D of NED, with or w/o DAAS

Through empirical analysis, we observed that visual disparities are sensitive to regions. We employed DAAS to direct the ASCCL's attention toward those challenging regions, thereby enhancing the consistency of mouth movements between input and output. In this section, We examined the influence of incorporating DAAS into the outcome. As depicted in Table 6, in the absence of DAAS, our approach outperforms NED across all metrics. Moreover, when DAAS is integrated into our method, ASCCL achieves even more optimal results by dynamically adapting the learning strategy to different regions.

6. Conclusion

In this work, we propose an adaptive spatial coherent correlation learning (ASCCL) algorithm to investigate the inherent visual correlations between the SPFEM model's input and output to construct paired supervision to improve facial expression manipulation and meanwhile better preserve the facial animation of speech content. It first characterizes the visual disparities, and then constrains the visual disparities in adjacent local regions of the input image to align with those in corresponding adjacent regions of the output image from the SPFEM model, together with a disparity-aware adaptive strategy to adaptively learn based on the visual disparity of each adjacent region. ASCCL is implemented by a plug-and-play loss that can be seamlessly integrated into any advanced methods to facilitate SPFEM performance. We conduct extensive experiments that integrate ASCCL to two advanced SPFEM models and carry out variant quantitative and qualitative comparisons as well as user studies to demonstrate the effectiveness.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

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

[5] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[6] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

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

[8] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[9] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023.

[10] Michail Christos Doukas, Mohammad Rami Koujan, Viktoriia Sharmanska, Anastasios Roussos, and Stefanos Zafeiriou. Head2head++: Deep facial attributes re-targeting. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):31–43, 2021.

[11] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[12] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Spectre: Visual speech-informed perceptual 3d facial expression reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5744–5754, 2023.

[13] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

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

[18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34: 852–863, 2021.

[19] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[21] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[23] Foivos Paraperas Papantoniou, Panagiotis P Filntisis, Petros Maragos, and Anastasios Roussos. Neural

emotion director: Speech-preserving semantic control of facial expressions in" in-the-wild" videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18781–18790, 2022.

[24] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020.

[25] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: One-shot anatomically consistent facial animation. *International Journal of Computer Vision*, 128:698–713, 2020.

[26] Girish Kumar Solanki and Anastasios Roussos. Deep semantic manipulation of facial videos. In *European Conference on Computer Vision*, pages 104–120. Springer, 2022.

[27] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020.

[28] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3385–3394, 2020.

[29] Rotem Tzaban, Ron Mokady, Rinon Gal, Amit Bermano, and Daniel Cohen-Or. Stitch it in time: Gan-based facial editing of real videos. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

[30] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020.

[31] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5021–5030, 2020.

[32] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. Ccpl: contrastive coherence preserving loss for versatile style transfer. In *European Conference on Computer Vision*, pages 189–206. Springer, 2022.

[33] Yongzong Xu, Zhijing Yang, Tianshui Chen, Kai Li, and Chunmei Qing. Progressive transformer machine for natural character reenactment. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s):1–22, 2023.

[34] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022.

[35] Bingyuan Zhang, Xulong Zhang, Ning Cheng, Jun Yu, Jing Xiao, and Jianzong Wang. Emotalker: Emotionally editable talking face generation via diffusion model. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2024.

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.