Learning Adaptive Spatial Coherent Correlations for Speech-Preserving Facial Expression Manipulation (Supplementary Material)

Due to the page limit in the main manuscript, we present some implementation details and experiment results and analyses for better reproducibility and completeness in this supplementary material. It consists of the following three aspects: 1) We present the network details of the ASCCL algorithms for better reproducibility. 2) We integrated ASCCL into two other SPFEM models [2, 5] and conducted experiments on the MEAD dataset. 3) We illustrate more visualization results and analyses using NED and ICface baselines on both the MEAD and RAVDESS datasets. 4) We report the user studies using the ICface baseline on MEAD and using NED and ICface baselines on RAVDESS. We've also added video comparison results to our supplemental materials.

1. Implementation Details

1.1. Constructing Paired Data

We utilize the MEAD dataset as the foundation for training our ASCCL algorithm. Specifically, we curate a subset of 7,650 video recordings featuring 36 distinct speakers from the MEAD dataset for the purpose of ASCCL algorithm training. Despite the presence of videos within the MEAD that feature a speaker uttering the same sentence in diverse emotional states, acquiring pairs of image data where an image of a sentence spoken in one emotional state corresponds to another image of the same sentence spoken in a different emotional state remains challenging. To address this, we employ the Dynamic Time Warping (DTW [1]) algorithm to align the Mel spectra of the two videos, thereby obtaining one-to-one training data. This paired data is then utilized to train the ASCCL algorithm.

1.2. Training details

In our proposed method of ASCCL, we employ ArcFace [3] as a feature extractor to discern multi-scale features of an image. By fine-tuning ArcFace with paired data, we can establish spatially coherent correlations within the feature space between two images. Specifically, we designate the image with the neutral emotion as x and the image with the alternate emotion as y. Utilizing ArcFace, we extract the features corresponding to x, denoted as x^{fl}, and similarly, features corresponding to y are denoted as y^{fl}. Here, $l \in [1, 2, 3, 4]$ signifies the feature output of the l-th block of ArcFace. Subsequently, we sample the region and generate both positive and negative samples at the feature level. We will illustrate this process with an example of a single sampling process. Initially, we sample two adjacent regions in x and align the region indices to the y side, while concurrently selecting a separate region in y randomly to construct a negative sample. The visual disparity between two adjacent regions, i and j, is represented as $x^{fl}_{i\to j}$ for image x, and as $y^{fl}_{i\to j}$ for image y. Similarly, the visual disparity between regions i and k in image y is denoted as $y^{fl}_{i\to k}$. For each i, we capture eight adjacent regions around it and denote this set j, while k represents a set of eight randomly sampled regions. Consequently, the pairings of $x^{fl}_{i\to j}$ and $y^{fl}_{i\to j}$ constitute positive samples, while the pairings of $x^{fl}_{i\to j}$ and $y^{fl}_{i\to k}$ form the negative samples. ASCCL aligns the visual consistency of two images by maximizing the similarity between positive samples and simultaneously distancing negative samples from each other. During the training phase, all images are resized with 224 × 224 resolution. We use Adam as an optimizer and set the batch size to 16. The initial learning rate is set to 1×10^{-4}. We conducted training for 50 epochs on the ASCCL using paired data, leveraging a single GeForce RTX 4090 graphics card. This process spanned approximately 30 hours.

Once ASCCL is trained, we fix its parameters and seamlessly integrate it into the SPFEM model. This integration serves as a guide during the training phase of the SPFEM model. The integration of ASCCL with the two-stage NED model begins with the first stage, where the inputs and outputs are 3DMM coefficients. In this stage, a 3D mesh is generated based on these coefficients, and the corresponding 2D images are created using the camera parameters specified by the application. ASCCL achieves visual consistency between the inputs and outputs by aligning the visual disparity between the corresponding adjacent regions of the two images. In the second stage, where both the input and output of the model are images, ASCCL can focus di-

rectly on enhancing the visual connection between the two images, thus supervising the training of the model more efficiently. When applied to ICface, a single-stage model in which both inputs and outputs are images, ASCCL aligns seamlessly with this structure. It guides model training by ensuring a high degree of visual consistency between similar inputs and outputs.

2. Supplementary Experiments

2.1. Quantitative Comparisons

We integrated ASCCL into these two SPFEM models [2, 5] and conducted experiments on the MEAD dataset [6], as shown in the Tables 1 and 2. GANmut [2] does not involve driven images, so there is no cross-ID and inter-ID setting. DSM [5] can extract the VA vector from the driven image, so we extracted the VA vector from reference videos and conducted tests under both cross-ID and inter-ID settings. As detailed in Tables 1 and 2, it also shows significant improvements when integrating ASCCL.

Emotions	GANmut			Ours (GANmut)		
	FID↓	LSE-D↓	CSIM↑	FID↓	LSE-D↓	CSIM↑
Neutral	3.450	8.965	0.806	2.888	8.866	0.819
Angry	7.203	8.961	0.662	6.853	8.862	0.659
Disgusted	8.687	8.959	0.643	8.741	8.861	0.621
Fear	7.953	8.979	0.592	7.512	8.863	0.606
Happy	6.667	8.957	0.613	6.248	8.857	0.629
Sad	8.335	8.968	0.601	8.138	8.868	0.599
Surprised	6.865	8.986	0.583	6.380	8.868	0.604
Avg.	7.023	8.968	0.643	6.680	8.864	0.648

Table 1. Comparision results of FAD, CSIM, and LSE-D of GANmut[2] with and without our ASCCL on the MEAD dataset

Settings	Emotions	DSM			Ours (DSM)		
		FAD↓	LSE-D↓	CSIM↑	FAD↓	LSE-D↓	CSIM↑
Inter-ID	Neutral	2.572	9.452	0.806	1.476	9.342	0.870
	Angry	2.156	9.835	0.780	2.183	9.656	0.880
	Disgusted	2.125	9.272	0.815	2.033	9.108	0.878
	Fear	2.364	9.676	0.790	2.092	9.695	0.887
	Happy	1.951	9.664	0.815	1.753	9.563	0.916
	Sad	1.985	9.594	0.821	1.794	9.487	0.917
	Surprised	1.908	9.226	0.818	1.794	9.131	0.916
	Avg.	2.152	9.531	0.806	1.875	9.426	0.895
Cross-ID	Neutral	1.916	9.801	0.866	2.008	9.409	0.877
	Angry	5.071	9.888	0.753	4.955	9.483	0.755
	Disgusted	4.991	10.157	0.784	4.976	9.364	0.785
	Fear	4.686	9.739	0.737	4.794	9.37	0.726
	Happy	5.274	9.518	0.777	4.447	9.482	0.880
	Sad	4.943	9.961	0.744	4.706	9.413	0.737
	Surprised	4.338	10.357	0.787	4.213	9.301	0.769
	Avg.	4.460	9.917	0.778	4.300	9.403	0.790

Table 2. Comparision results of FAD, CSIM, and LSE-D of DSM [5] with and without our ASCCL on the inter-IDentification and cross-IDentification settings on the MEAD dataset

2.2. Qualitative Comparisons

In the main manuscript, we present visualization examples using NED with and without the proposed ASCCL algorithms on the MEAD dataset. Here, we further supplement more examples of the NED and ICface with and without the proposed ASCCL algorithms on both MEAD [6] and RAVDESS [4] datasets.

The visualization results using NED and ICface baseline on the MEAD dataset are presented in Figure 1 and 3. We have presented detailed analyses of using the NED baseline in the main manuscript. Here, we mainly analyze the qualitative comparisons of using ICface baseline in detail for the following three aspects. 1) Realism. ICface executes the SPFEM task by mapping the corresponding action units from the reference image to the source image. This procedure often muddles the identity information of the result and reference images, as the action units are not wholly uncoupled from the reference image's identity information, as illustrated in the first and fourth rows of the third column in Figure 3. ASCCL mitigates this issue to some extent by harmonizing the visual consistency between the source and result images, as depicted in the fourth column of Figure 3. 2) Emotion similarity. ICface utilizes action units that are not fully decoupled to accomplish the SPFEM task, resulting in ID confusion or even image distortion as shown in the third column of the first, third, and fifth rows of Figure 3, which leads to low or even unrecognizable emoticon similarity between source and result. ASCCL accomplishes the emotion migration by aligning the source and result corresponding to the visual disparity in adjacent regions, which to some extent preserves the invariant information between the two to accomplish the emotion migration, as shown in the fourth column of Figure 3. 3) Lip-audio preserving accuracy. Since the action units are not completely decoupled from the image ID information, when exchanging action units between source and reference, the ICface cannot realize the preservation of the mouth shape, and even the reference image will be directly copied to the source side as shown in the third column of Figure 3. The integration of ASCCL into the ICface empowers the ICface to preserve the visual consistency between the inputs and the outputs, and to a certain extent maintains the mouth shape while preserving the ID information unchanged as shown in the fourth column of Figure 3

Similarly, we present the results using NED and ICface baseline on the RAVDESS dataset in Figures 2 and 4. We can observe the improvement in all three aspects of realism, emotion similarity, and lip-audio preserving accuracy are similar to those on the MEAD dataset. These comparisons also demonstrate that the proposed ASCCL can generalize to different datasets.

2.3. User Study

In the main manuscript, we report the user study results using NED with and without the proposed ASCCL algorithms on the MEAD dataset. Here, we further supplement more user study results of the NED and ICface with and with-

Emotion	Realism		Emotion similarity		Mouth shape similarity	
	ICface	ASCCL	ICface	ASCCL	ICface	ASCCL
Neutral	39%	61%	42%	58%	42%	58%
Angry	39%	61%	31%	69%	39%	61%
Disgusted	50%	50%	39%	61%	31%	69%
Fear	31%	69%	39%	61%	28%	72%
Happy	33%	67%	36%	64%	47%	53%
Sad	33%	67%	31%	69%	33%	67%
Surprised	47%	53%	39%	61%	44%	56%
Avg.	39%	61%	37%	63%	38%	62%

Table 3. Realism, emotion similarity, and mouth shape similarity ratings of the user study for ICface and our ASCCL on the MEAD dataset.

Emotion	Realism		Emotion similarity		Mouth shape similarity	
	NED	ASCCL	NED	ASCCL	NED	ASCCL
Neutral	29%	71%	21%	79%	29%	71%
Angry	42%	58%	29%	71%	46%	54%
Disgusted	58%	42%	46%	54%	54%	46%
Fear	42%	58%	42%	58%	46%	54%
Happy	38%	62%	33%	67%	50%	50%
Sad	38%	62%	21%	79%	46%	54%
Surprised	38%	62%	38%	62%	46%	54%
Avg.	40%	60%	33%	67%	45%	55%

Table 4. Realism, emotion similarity, and mouth shape similarity ratings of the user study for NED and our ASCCL on RAVDESS dataset.

out the proposed ASCCL algorithms on both MEAD [6] and RAVDESS [4] datasets. We have presented detailed analyses of using the NED baseline in the main manuscript. Here, we mainly analyze the qualitative comparisons of using the ICface baseline on the MEAD dataset. The setting is the same as that we described in the main manuscript: 10 videos per emotion, 70 videos in total, and 25 participants. As shown in Table 3, incorporating the ASCCL algorithm can significantly improve the realism, emotion similarity, and lip synchronicity, achieving several times more rate than the ICface baseline across all seven emotions. On average, incorporating the ASCCL algorithm obtains 22% more rate on realism, 26% more rate on emotion similarity, and 24% more rate on mouth shape similarity compared with the NED baseline. Similarly, we present the results using NED and ICface baseline on the RAVDESS dataset in Table 4 and Table 5. As RAVDESS contains fewer videos, we randomly select 5 videos per emotion, resulting in a total of 35 videos. Then, we also asked 25 participants for evaluation. We find incorporating ASCCL also obtains very obviously more rates in all three aspects.

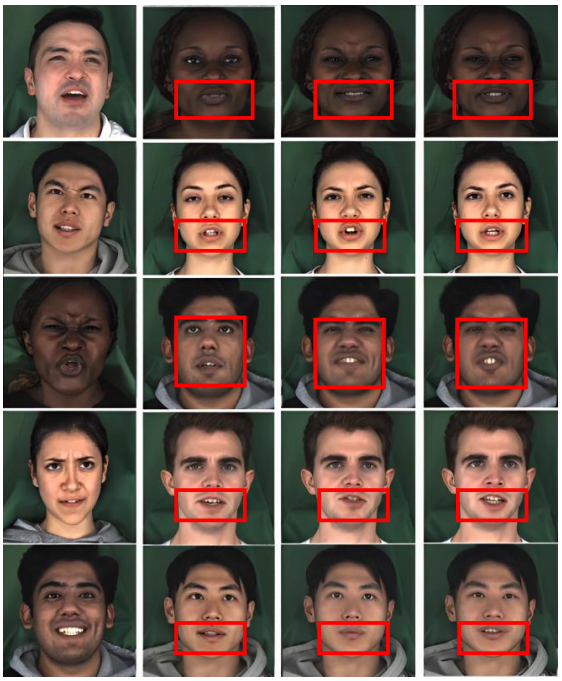

Figure 1. Qualitative comparisons of NED with and without the proposed ASCCL algorithm. The samples are selected from the MEAD dataset.

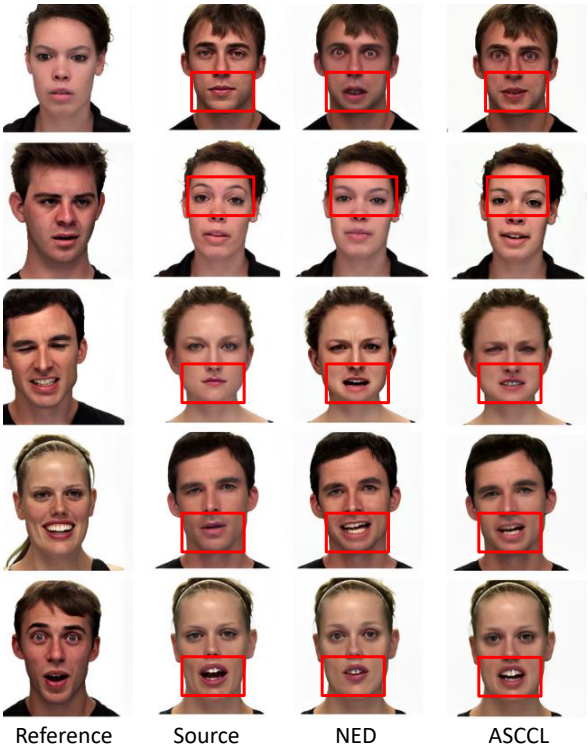

Figure 2. Qualitative comparisons of NED with and without the proposed ASCCL algorithm. The samples are selected from the RAVDESS dataset.

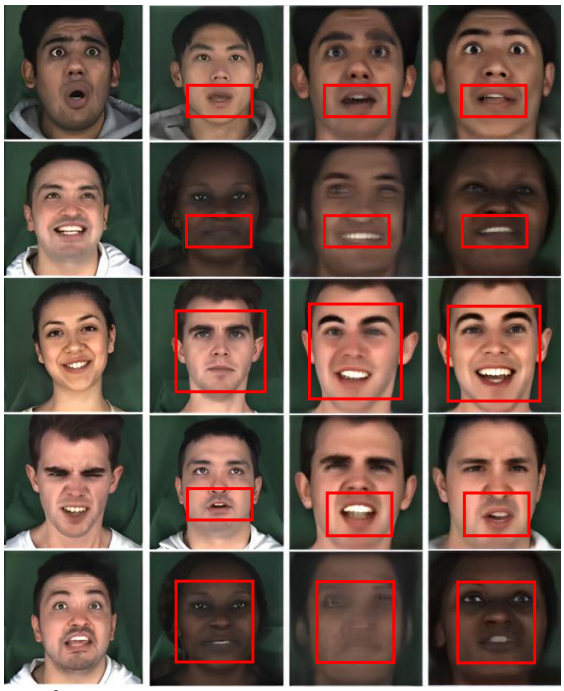

Figure 3. . Qualitative comparisons of ICface with and without the proposed ASCCL algorithm. The samples are selected from the MEAD dataset.

Emotion	Realism		Emotion similarity		Mouth shape similarity	
	ICface	ASCCL	ICface	ASCCL	ICface	ASCCL
Neutral	22%	78%	24%	76%	31%	69%
Angry	28%	72%	36%	64%	30%	70%
Disgusted	20%	80%	34%	66%	26%	74%
Fear	28%	72%	35%	65%	29%	71%
Happy	22%	78%	33%	67%	28%	72%
Sad	25%	75%	28%	72%	25%	75%
Surprised	20%	80%	27%	73%	24%	76%
Avg.	24%	76%	31%	69%	27%	73%

Table 5. Realism, emotion similarity, and mouth shape similarity ratings of the user study for ICface and our ASCCL on RAVDESS dataset.

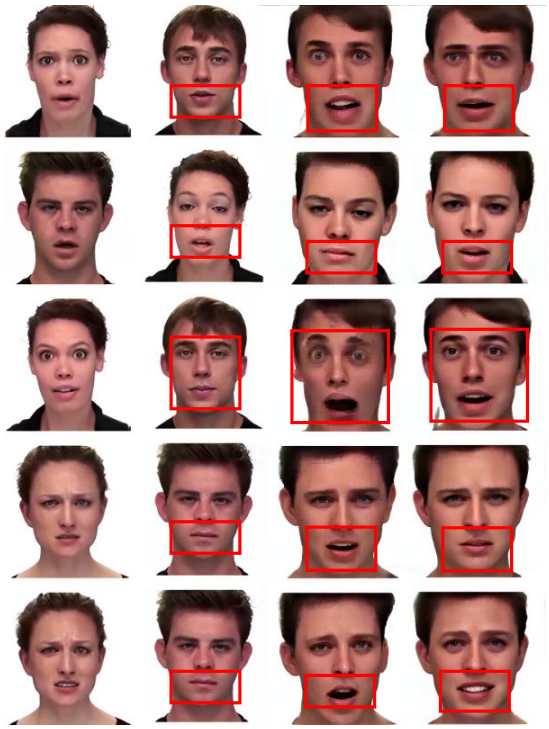

Figure 4. Qualitative comparisons of ICface with and without the proposed ASCCL algorithm. The samples are selected from the RAVDESS dataset.

References

[1] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, pages 359–370. Seattle, WA, USA:, 1994.

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

[3] 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.

[4] 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.

[5] Girish Kumar Solanki and Anastasios Roussos. Deep semantic manipulation of facial videos. In *European Conference on Computer Vision*, pages 104–120. Springer, 2023.

[6] 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.