Supplementary of "Texture, Shape and Order Matter: A New Transformer Design for Sequential DeepFake Detection"

1. Further Analysis

Effect of using Different Offsets in Attention Heads. Section 3.2.2 introduces the shape-guided Gaussian mapping combined with multi-head cross-attention. This process generates an independent head-specific offset $\Delta\boldsymbol{\mu}_i$ and $\Delta\boldsymbol{\sigma}_i$ for each attention head. To verify the effect of offset, we remove this offset and make all attention heads share the same mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$. The results in Table 1 indicate a performance drop of 0.51% in FACC and 0.92% in ACC without the offset highlighting the importance of the head-specific offset.

Table 1. Effect of using offsets in attention heads.

	FACC	AACC
w/o offset	74.70	58.82
w/ offset	75.21 (+0.51)	59.74 (+0.92)

Effect of Adaptive Coefficient Vector $\boldsymbol{\alpha}$. This part studies the effect of using adaptive coefficient vector $\boldsymbol{\alpha}$ (refer to Section 3.1.2). Our method adaptively balances the attention to central, angular, and radial difference operations, with $\boldsymbol{\alpha} = \mathrm{Softmax}(\mathcal{H}(\mathbf{x}'))$. To demonstrate its effectiveness, we manually set fixed values $\boldsymbol{\alpha} = \{1/3, 1/3, 1/3\}$, equalizing the importance of the three convolutional differential operations. As shown in Table 2, the adaptive $\boldsymbol{\alpha}$ with the learnable coefficients achieves higher accuracy than the fixed weight coefficients, as expected.

Table 2. Effect of adaptive coefficient vector $\boldsymbol{\alpha}$.

	FACC	AACC
$\boldsymbol{\alpha} = \{1/3, 1/3, 1/3\}$	74.76	58.74
$\boldsymbol{\alpha} = \mathrm{Softmax}(\mathcal{H}(\mathbf{x}'))$	75.21 (+0.45)	59.74 (+1.00)

Further Exploration of Prediction Order. This part further discusses the effect of using various prediction orders. As studied in the main text, the inverted order prediction performs better than the regular forward order. However, the performance of using mixed order has not been explored. Note that the annotation length for facial manipulation is five. Besides the complete forward order (5FO) and inverted order (5IO), we also study a set of mixed orders, which includes three forward orders and two inverted orders (3FO, 2IO), and two forward orders and three inverted orders (2FO, 3IO). The results are shown in Table 3, indicating that the fully inverted order performs best.

Table 3. Effect of various orders.

Methods	FACC	AACC
5 FO, 0 IO	71.85	54.26
3 FO, 2 IO	71.73	53.59
2 FO, 3 IO	72.76	55.89
0 FO, 5 IO	75.21	59.74

Different Stem Architectures. We conduct additional experiments using different stem architectures. As shown in Table 4, our method improves alongside the improvement of the network capabilities, achieving the best results with ResNet-101. This fully aligns with our expectation that a stronger stem can further improve the ability to capture tampering features.

Table 4. Performance of stem architectures.

	FACC	AACC
Resnet-18	72.96	56.09
Resnet-34	75.21	59.74
Resnet-50	75.53	59.67
Resnet-101	75.59	60.25

Limitations. Since our method is designed specifically for sequential DeepFake detection, it shares a limitation with existing methods in its capacity for one-step DeepFake detection. In future work, we aim to explore solutions that can improve generalizability across both one-step and sequential DeepFake detection.

1.1. More Details of Reproduction

We reproduce the methods of DRN, MA, Two-stream, and SeqFake-Former using their official codes and rigorously follow the training instructions. Note that the reproduced results of DRN, MA, and Two-stream are consistent with, even better than those reported in Paper [1]. However, the results of SeqFake-Former are slightly lower than its original report. This is because these methods are originally executed on four GPUs, whereas we only use a single GPU in reproduction due to limited computing resources. This difference can restrict the usage of large batch sizes, leading to a certain performance drop.

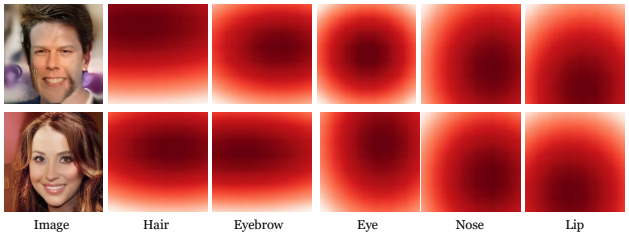

Figure 1. Shape priors visualization.

2. More Visualizations

Shape Priors Visualization. We visualize the probability maps S in Fig. 1 to assess whether shape priors are well learned. Note the manipulation annotations of these two images are eyebrow-eye-hair-nose-lip and nose-eyebrow-lip-eye-hair, respectively. It can be seen that the corresponding probability map can roughly reflect the manipulation regions.

Examples of Challenging Scenarios. Fig.2 shows examples of images applying post-processing operations, including Gaussnoise, Image compression, ColorJitter, ToGray, and RGBShift (implemented using **albumentations** API).

The parameters for each operation are as follows: For Gaussian noise, we set the var-limit range from 10 to 50, with a mean of zero, and independently sample the noise for each channel. For the image compression operation, the quality bounds are set from 25 to 50. For jittering colors, we use a contrast and brightness setting of 0.7 to 1.4, while keeping the saturation unchanged. For the RGB shift operation, we set the shift-limit values for the three channels to range from -20 to 20.

Figure 2. Examples of challenging scenarios.

More Attention Visualizations. Fig 3 presents more attention visualization results of our method. It can be seen that our model can capture the manipulated regions with various lengths of manipulation annotations, including Eye, Eye-Nose, Eyebrow, Eye-Nose-Lip-Eyebrow respectively. Notably, our method can maintain focused attention on one manipulation without affecting the attention of other manipulated attributes.

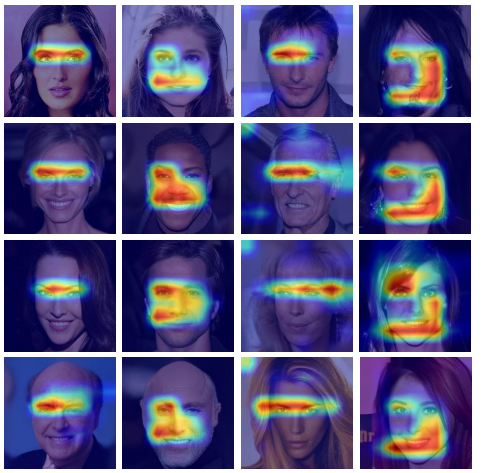

Figure 3. More attention visualizations of different manipulating annotations.

References

[1] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and recovering sequential deepfake manipulation. In *European Conference on Computer Vision (ECCV)*, 2022.