# Supplementary Material for *Exploring Frequency Attention Learning and Contrastive Learning for Face Forgery Detection*

Neng Fang, Bo Xiao(✉), Bo Wang, Chong Li, and Lanxiang Zhou

Beijing University of Posts and Telecommunications
{fangneng,xiaobo,bobo,lch1203,zhoulanxiang}@bupt.edu.cn

**Abstract.** This material shows more experimental results, visualization, and analysis for our Frequency Attention Learning and Contrastive Learning (FACL) due to space limitation of the main paper.

## 1 Explanation of DFCL

To make full understanding of our DFCL, we will discuss the relationship among our $\mathcal{L}_{DFCL}$ with the triplet loss [3]. Specifically, let $R^a$ and $F^a$ to represent the real and fake anchors, and $R^p$ and $F^p$ to represent the real and fake positives, respectively. We can demonstrate that the triplet loss is a special case of our DFCL loss:

$$\begin{aligned}
\mathcal{L}_{DFCL} &= -\log \frac{e^{\delta(R^a,R^p)/\tau}}{e^{\delta(R^a,R^p)/\tau}+e^{\delta(R^a,F^a)/\tau}+e^{\delta(R^a,F^p)/\tau}} \\
&= \log\left(1+\frac{e^{\delta(R^a,F^a)/\tau}+e^{\delta(R^a,F^p)/\tau}}{e^{\delta(R^a,R^p)/\tau}}\right) \\
&\doteq e^{\delta(R^a,F^a)/\tau-\delta(R^a,R^p)/\tau}+e^{\delta(R^a,F^p)/\tau-\delta(R^a,R^p)/\tau} \\
&\doteq 2+\delta\left(R^a,F^a\right)/\tau+\delta\left(R^a,F^p\right)/\tau-2\delta\left(R^a,R^p\right)/\tau \\
&= 2+\frac{1}{2\tau}\left(2\left\|R^a-R^p\right\|_2^2-\left\|R^a-F^a\right\|_2^2-\left\|R^a-F^p\right\|_2^2\right) \\
&\propto 2\left\|R^a-R^p\right\|_2^2-\left\|R^a-F^a\right\|_2^2-\left\|R^a-F^p\right\|_2^2+4\tau
\end{aligned} \tag{1}$$

which bears a resemblance to the triplet loss, but incorporates quadruple variables. In contrastive learning, hard mining is an essential aspect for achieving optimal performance with the triplet loss and other metric learning losses [2]. It is worth noting that we did not pull the manipulated faces closer in the feature space, as we recognized the inherent distinctions between various forgery techniques.

## 2 A plug-and-play Frequency Attention Module

In Section 3.2 of the main paper, we present a novel method to automatically extract discriminative frequency features through our frequency attention module

(FAM). Our FAM can be seamlessly integrated with various spatial-aware face forgery detection techniques, significantly improving their ability to distinguish between real and fake faces. Here, we replace the visual encoder in Fig. 2 with commonly-used visual backbones (*e.g.*, ResNet [1], EfficientNet [4]), resulting in **Base** which operates solely in the spatial domain and **Base w/ FAM** which incorporates our FAM. The quantitative results are shown in Table 1. We observe that our FAM consistently help improve the performance of the **Base** model. It is worth nothing that our FAM achieves a much higher average improvement on low quality (C40) videos than on high quality (C23) videos. This indicates that our FAM enables the model to effectively identify discriminative artifact clues in the frequency domain, which is especially critical under the low quality setting. To conclude, our FAM is a plug-and-play module that enhances the learning ability of frequency-based clues for detecting face forgery.

**Table 1.** Ablations of different visual backbones with our frequency attention module (FAM). Red results indicate the best.

| **Compression** | **Model** | **Backbone** | | |
|---|---|---|---|---|
| | | **Xception** | **ResNet50** | **Efficientnet-b4** |
| C23 | Base | 93.31% | 94.67% | 99.18% |
| | Base w/ FAM | 99.15% | 97.21% | 99.63% |
| C40 | Base | 81.76% | 81.83% | 88.20% |
| | Base w/ FAM | 93.76% | 93.22% | 94.03% |

## 3 Visualizations of reconstruction module

As illustrated in Table 2, **Base w/ REC** improved the the AUC by 0.7% (91.97→92.67) from **Base**. The improvement is attributed to the ground-truth mask guides the model's focus towards the tampered regions during feature extraction. Additionally, we visualize the decoder output in *supp*, which provides qualitative evidence of the effectiveness of our reconstruction module in restoring tampered regions. The results presented in Table 2 demonstrate the effectiveness of our reconstruction module, as it improved the AUC by 0.7% (91.97→92.67) from **Base** to **Base w/ REC**. This improvement is attributed to the binary ground-truth mask that guides the model's focus towards the tampered regions during feature extraction. Additionally, we visualize the decoder output with the ground-truth forgery mask in Fig.1, which displays the reconstructed results of the decoder in our FACL. We observe the predicted areas verge on the ground-truths, accurately capturing the local forged regions generated by different face manipulated algorithms. Meanwhile, reconstructions with different manipulated techniques vary significantly, indicating that our FACL possesses well robustness in the cross-manipulation scenario. For genuine faces, the predicted areas are almost entirely black, further demonstrating the efficacy of our FACL in

**Table 2.** Ablations of the impact of our proposed components. RGB: the RGB stream input, FAM: Frequency Attention Module, DFCL: DeepFake Contrasitive Loss, REC: the decoder with the reconstruction task. Red numbers indicate the best performance.

| Method | RGB | FAM | DFCL | REC | AUC |
|---|---|---|---|---|---|
| Base | ✓ | | | | 91.97% |
| Base w/ FAM | ✓ | ✓ | | | 93.76% |
| Base w/ DFCL | ✓ | | ✓ | | 93.44% |
| Base w/ REC | ✓ | | | ✓ | 92.67% |
| FACL | ✓ | ✓ | ✓ | ✓ | **94.33%** |

predicting modified pixels for forged faces and yielding superior performance. provides qualitative evidence of the effectiveness of our reconstruction module in restoring tampered regions.

Moreover, our proposed Frequency Attention Module, Deepfake Contrastive Loss, and the reconstruction decoder contribute to the performance improvement of our full FACL. As shown in Table 2, our **FACL** outperforms **Base w/ FAM** by 0.6% (93.76→94.33), **Base w/ DFCL** by 0.9% (93.44→94.33), and **Base w/ REC** by 1.8% (92.67→94.33). In conclusion, our FACL consistently improves the performance by incorporating our novel modules and the multi-task two-stream architecture.

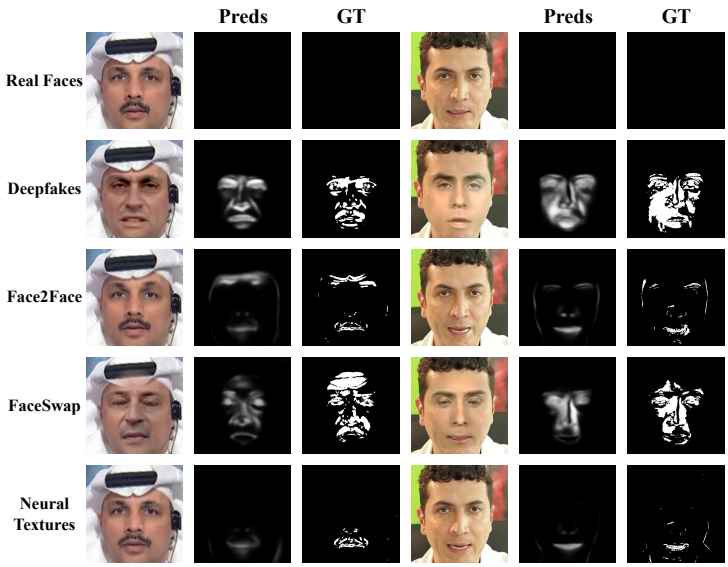


**Fig. 1.** The visualization of the predicted manipulated areas (Preds) and the ground truth (GT).

## References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
2. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: NIPS (2020)
3. Kumar, A., Bhavsar, A., Verma, R.: Detecting deepfakes with metric learning. In: IWBF (2020)
4. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019)