Decoupling Forgery Semantics for Generalizable Deepfake Detection: Supplementary Material

Wei Ye
weiye@email.ncu.edu.cn

Xinan He
shahur@email.ncu.edu.cn

Feng Ding*
fengding@ncu.edu.cn

Nanchang University
Nanchang, China

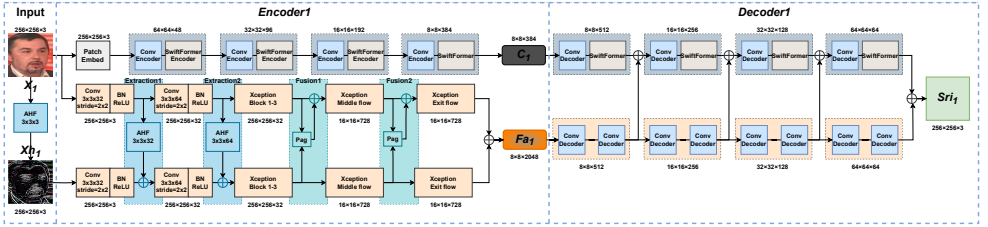## 1   The Network Details of Training Stage 1

Figure 1: The architecture details of $Encoder_1$ and $Decoder_1$ in our proposed method. (**Left**) In $Encoder_1$, $Extraction_1$ and $Extraction_2$ constitute a multi-scale high-frequency feature extraction module (MHFE), while $Fusion_1$ and $Fusion_2$ form a multi-scale high-frequency feature fusion module (MHFF). (**Right**) In $Decoder_1$, convolutional layers and Swiftformer are used to reconstruct images.

Encoder1. We use SwiftFormer-L1 [2] to extract irrelevant content from images and employ Xception [1] for extracting RGB information and high-frequency features. Figure 1 illustrates the process for inputs X_1 and Xh_1, obtaining irrelevant content semantics C_1 and all forgery semantics Fa_1. Consistently, for inputs X_0 and Xh_0, we obtain irrelevant content semantics C_0 and all forgery semantics Fa_0.

Decoder1. In $Decoder_1$, we designed a dual-channel network. One channel simultaneously uses convolutional layers and SwiftFormer [2] to process irrelevant content semantics, while the other channel solely uses convolutional layers to process all forgery semantics. Figure 1 illustrates the process for inputs C_1 and Fa_1, resulting in the self-reconstructed image Sri_1. Similarly, inputs C_0 and Fa_0 yield the self-reconstructed image Sri_0. Additionally, inputs C_1

---

*Corresponding author

and Fa_0 produce the cross-reconstructed image Cri_1, while inputs C_0 and Fa_1 generate the cross-reconstructed image Cri_0.

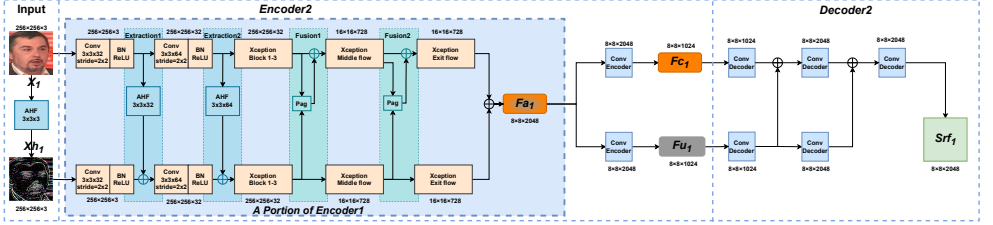# 2 The Network Details of Training Stage 2

Figure 2: The architecture details of $Encoder_2$ and $Decoder_2$ in our proposed method. (**Left**) In $Encoder_2$ of training stage 2, a branch first utilizes $Encoder_1$ to extract all forgery semantics for extracting common forgery semantics. (**Right**) In $Decoder_2$, both branches solely employ convolutional layers to reconstruct forgery semantics.

Encoder2. In $Encoder_2$, we utilize a portion of $Encoder_1$ responsible for extracting all forgery semantics to extract all forgery semantics, and then employ convolutional layers further disentangle these semantics into unique and common forgery semantics. Figure 2 illustrates the process for inputs X_1 and Xh_1, obtaining unique forgery semantics Fu_1 and common forgery semantics Fc_1. Similarly, for inputs X_0 and Xh_0, we obtain unique forgery semantics Fu_0 and common forgery semantics Fc_0.

Decoder2. In $Decoder_2$, we designed two dual-channel networks, each comprising only convolutional decoders, and merged them during the process to reconstruct image semantics. Figure 2 illustrates the process for input Fc_1 and Fu_1 to obtain self-reconstructed image semantics Srf_1. Consistently, for inputs Fc_0 and Fu_0, we obtain self-reconstructed image semantics Srf_0. Additionally, inputs Fc_1 and Fu_0 yield cross-reconstructed image semantics Crf_1, while inputs Fc_0 and Fu_1 yield cross-reconstructed image semantics Crf_0.

References

[1] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[2] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17425–17436, 2023.