An End-to-End Framework for Joint Deepfake Detection and Fine-Grained Localization

Haotian Liu[1], **Chenhui Pan**[2], **Ying Chen**[2], **Changfa Mo**[2], **Guoying Zhao**[1] and **Xiaobai Li**[2,1*]

[1]Center for Machine Vision and Signal Analysis, University of Oulu, Finland

[2]State Key Laboratory of Blockchain and Data Security, Zhejiang University, China

Abstract

Deepfake detection focuses on identifying forged content in visual data. With the rapid advances in generative models, it has become increasingly critical to mitigate risks such as identity fraud, misinformation, and other security threats. Most existing Deepfake detection methods primarily focus on image-level classification. Despite their notable progress, they generally lack the capability to precisely localize the forgery regions, which limits their applicability and interpretability in real-world settings, especially in scenarios involving multiple faces or facial component manipulations. To this end, we propose a unified end-to-end framework that jointly performs image-level forgery classification and fine-grained localization of forgery regions. Considering that forgery regions are often small and sparsely distributed, we utilize a set of learnable queries and masked attention mechanisms to suppress background noise and guide the localization prediction to focus on relevant regions. Extensive experiments on the DDL-I challenge benchmark validate the effectiveness of our framework for both Deepfake detection and localization tasks. By using the ensemble strategy, our final solution achieves an overall score of 81.50% on the DDL-I testing set, demonstrating strong performance in handling practical Deepfake detection challenges.

1 Introduction

Deepfake detection [Juefei-Xu *et al.*, 2022] aims to identify manipulated content in visual data. With the rapid advances in deep generative models, producing highly realistic synthetic facial images has become increasingly accessible. However, the misuse of AI-generated content presents serious risks, including identity fraud, misinformation, and online scams [Lin *et al.*, 2024]. Therefore, developing effective and interpretable Deepfake detection methods is essential to uphold digital authenticity and social trust.

*Corresponding Author. Email: xiaobai.li@zju.edu.cn.

Most existing Deepfake detection methods, e.g., [Rossler *et al.*, 2019], focus on image-level classification, aiming to determine whether an entire image is real or fake. These approaches leverage various artifact cues, such as spatial inconsistency [Chai *et al.*, 2020] and frequency information [Li *et al.*, 2021; Miao *et al.*, 2023b]. Despite notable progress, they often struggle to provide fine-grained localization of manipulated regions, which limits their effectiveness in real-world applications. Although some tailored Deepfake localization approaches, such as [Kong *et al.*, 2022; Shuai *et al.*, 2023; She *et al.*, 2024], have also been explored, they still fall short in real-world scenarios, particularly when dealing with high-resolution images that contain both genuine and forged faces or involve only subtle manipulations of specific facial components.

To address these challenges, we propose an end-to-end framework that simultaneously performs image-level forgery classification and precise localization of manipulated regions. Our model takes the entire image as input, supporting multi-face scenarios without pre-processing steps such as face cropping, thereby greatly simplifying the inference pipeline. Our method employs a shared backbone to extract features for both detection and localization tasks. Inspired by a modern Transformer-based segmentation model, Mask2Former [Cheng *et al.*, 2022], we utilize a set of learnable queries and masked attention mechanisms to suppress background noise and make the localization focus on small and sparse forgery regions. Mask2Former was originally designed for semantic segmentation, where object regions are large and have clear boundaries, while Deepfake localization requires detecting small, sparsely distributed forgery regions. Therefore, our framework leverages learnable queries and masked attention mechanism to progressively refine localization regions that contain subtle manipulation artifacts.

Extensive experiments on the Deepfake Detection and Localization image (DDL-I) [Miao *et al.*, 2025] challenge benchmark show that the proposed method can effectively perform Deepfake detection and localization jointly, achieving an AUC of 94.95% for detection and an IoU of 71.79% for localization on the DDL-I testing set. With the ensemble strategy, our final solution got an overall score of 81.50% on the DDL-I testing set, achieving a leading position in the *IJCAI 2025 Challenge "The Deepfake Detection and Localization"*. These results highlight its strong capability in de-

tecting various types of forgeries and its potential for complex and challenging real-world application scenarios.

2 Related Work

2.1 Deepfake Detection

Deepfake detection focuses on distinguishing between authentic and forged facial images. Early studies [Rossler *et al.*, 2019] utilized conventional networks to perform binary image classification. Subsequent methods also explored more specific spatial artifacts, including eye blinking [Haliassos *et al.*, 2021], local patch inconsistency [Chai *et al.*, 2020], and frequency domain clues [Li *et al.*, 2021; Miao *et al.*, 2023b]. With the rise of deep generative models, recent studies have focused on improving the generalization of Deepfake detectors to unseen forgery types. Shiohara et al. [Shiohara and Yamasaki, 2022] proposed a data synthesis method that makes models to learn generalizable representations from diverse pseudo-samples. CDDB [Li *et al.*, 2023] adopted a continual learning paradigm to accommodate new forgery types without catastrophic forgetting. Recent approaches [Ojha *et al.*, 2023; Tan *et al.*, 2023; Liu *et al.*, 2024; Nguyen *et al.*, 2024] also utilized intermediate features or gradient information as universal representations to improve cross-domain generalization. However, these methods focus solely on image-level classification and lack the ability to localize forged regions, limiting their interpretability and real-world applicability.

2.2 Deepfake Localization

Deepfake localization aims to precisely locate forgery regions and predict pixel-level masks. While typical image manipulation localization methods [Liu *et al.*, 2022; Zhou *et al.*, 2023] performed well in natural scene images, they are less effective for Deepfake localization due to high-level semantic changes to facial features or identities in Deepfakes. Recent studies have proposed various strategies tailored for Deepfake localization. Kong et al. [Kong *et al.*, 2022] fused semantic and noise maps to extract high-level clues indicative of forgery. In contrast, Shuai et al. [Shuai *et al.*, 2023] introduced a two-stream architecture that integrates spatial information with noise residual cues. Other approaches leveraged attention-based mechanisms [Dang *et al.*, 2020] and graph-based reasoning frameworks [She *et al.*, 2024] to improve spatial sensitivity and structural modeling. In parallel, more generalized manipulation localization methods [Ma *et al.*, 2023; Zhu *et al.*, 2025] capable of handling Deepfakes alongside other image manipulation types have also been explored. Besides, unified frameworks [Zhang *et al.*, 2024a; Miao *et al.*, 2024; Miao *et al.*, 2023a; Zhang *et al.*, 2024b] that tackle Deepfake detection and localization jointly have also been proposed.

Despite the swift progress, most existing methods overlook the small and sparse nature of facial forgeries, and they depend on pre-processing steps like face detection and cropping, limiting their localization performance in high-resolution images, especially in challenging real-world scenarios involving facial components or multiple face manipulations.

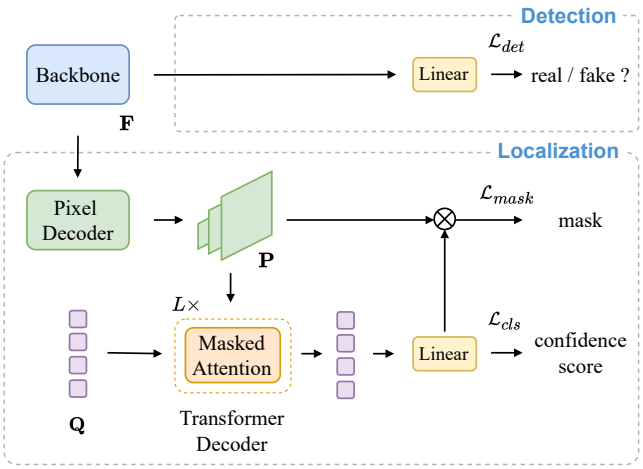

Figure 1: The overview of our proposed framework. It performs Deepfake detection and localization in an end-to-end manner. A shared backbone is used to extract feature maps $\mathbf{F}$. For localization, per-pixel embeddings $\mathbf{P}$ are generated from pixel decoder, and learnable queries $\mathbf{Q}$ are used to decode localization predictions.

3 Method

The overall framework of our proposed method is illustrated in Figure 1. We adopt an end-to-end architecture that jointly performs image-level Deepfake detection and fine-grained Deepfake localization. The model takes the entire image as input, potentially containing one or multiple faces. The shared backbone extracts feature maps, which are then used to predict classification scores as well as localization masks indicating forgery regions. This method enables efficient image processing without extra pre-processing steps like face detection and cropping, thereby simplifying the pipeline for unified Deepfake detection and localization.

3.1 Backbone

Deepfake detection and localization require a large receptive field to capture global manipulation cues, as well as fine-grained representations for precise localization of manipulated regions. To this end, we adopt Swin Transformer [Liu *et al.*, 2021] as the shared backbone for both tasks, which leverages shifted window attention to efficiently model long-range pixel-wise interactions. Given an image with resolution $H \times W$ as input, the backbone generates the feature map $\mathbf{F} \in \mathbb{R}^{C_F \times \frac{H}{S} \times \frac{W}{S}}$, where C_F is the channel and S is downsampling ratio.

3.2 Masked Deepfake Localization

Deepfake localization is challenging due to facial manipulation pixels are sparse relative to background regions, and a single image may contain both full-face swaps and facial component manipulation. To address this, we adopt Mask2Former [Cheng *et al.*, 2022], introducing learnable queries and utilizing a masked transformer decoder to efficiently locate manipulated regions.

Pixel Decoder. We follow Mask2Former and use a pixel decoder to upsample feature maps from backbone, obtaining a set of high-resolution per-pixel embeddings $\mathbf{P} \in \mathbb{R}^{C\times H\times W}$. In this process, deformable attention layers are employed to fuse information across feature maps.

Masked-Attention Transformer Decoder. The Transformer decoder introduces a set of learnable queries $\mathbf{Q} \in \mathbb{R}^{N\times C}$ to decode per-pixel embeddings into localization predictions. Specifically, it consists of L Transformer layers. Each layer performs cross-attention with per-pixel embeddings $\mathbf{P}$ to iteratively update the learnable queries and refine the localization mask prediction.

To enhance the decoder's focus on small and sparse manipulated regions, we utilize **masked cross-attention** [Cheng *et al.*, 2022], allowing each query to focus selectively on relevant regions. This process is formulated as:

$$\tilde{\mathbf{Q}} = \mathrm{softmax}\left(\mathbf{Q}_l\mathbf{K}_l^T + \mathcal{M}\right)\mathbf{V}_l + \mathbf{Q}. \tag{1}$$

Here, $\mathbf{Q}$ and $\tilde{\mathbf{Q}}$ denote the input and updated learnable queries, respectively. $\mathbf{Q}_l$ at l-th layer is projected from $\mathbf{Q}$, while $\mathbf{K}_l$ and $\mathbf{V}_l$ are obtained from per-pixel embeddings $\mathbf{P}$.

The attention mask $\mathcal{M}$ at position (i, j) is derived from the previous layer's mask prediction $\mathbf{M}_{l-1} \in [0, 1]^{N\times HW}$:

$$\mathcal{M}(i,j) = \begin{cases} 0, & \text{if } \mathbf{M}_{l-1}(i,j) > 0.5, \\ -\infty, & \text{otherwise.} \end{cases} \tag{2}$$

Localization Loss. Each learnable query is then decoded through linear layers to predict a pixel-wise localization mask and a foreground confidence score. The localization loss is defined as:

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{cls}}, \tag{3}$$

where $\mathcal{L}_{\text{mask}}$ combines pixel-wise binary cross-entropy and Dice losses for localization mask, while $\mathcal{L}_{\text{cls}}$ is a binary cross-entropy loss for confidence score predictions, as described in [Cheng *et al.*, 2022].

3.3 Deepfake Detection

To identify whether the entire image is real or fake, we directly feed the backbone feature map $\mathbf{F}$ into a lightweight linear layer. A binary cross-entropy loss for detection is computed as:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{bce}}(y, \hat{y}) \tag{4}$$

where $y \in \{0, 1\}$ is the ground truth label, and $\hat{y}$ is the predicted probability.

Note that the detection and localization branches are structurally independent, enabling efficient processing for either task individually or both jointly.

3.4 Optimization

The overall framework is end-to-end optimized by:

$$\mathcal{L} = \lambda_{det} \cdot \mathcal{L}_{\text{det}} + \lambda_{loc} \cdot \mathcal{L}_{\text{loc}}, \tag{5}$$

where λ_{det} and λ_{loc} denote the loss weights for Deepfake detection and localization, respectively. We adopt $\lambda_{det} = 0.1$ and $\lambda_{loc} = 1$ to achieve balanced convergence for both tasks.

3.5 Ensemble

Ensembling is a common strategy in deep learning to combine predictions from multiple models, aiming to improve overall performance. We adopt this approach to fuse the outputs of two models with different backbones.

The ensemble prediction Y is calculated as follows:

$$Y = \frac{w_1 \cdot y_1 + w_2 \cdot y_2}{w_1 + w_2}, \tag{6}$$

where y_1 and y_2 denote the outputs of the two models from either Deepfake detection or localization. The fusion weights w_1 and w_2 are used to balance the contributions of the two models.

4 Experiments

4.1 Experiment Setup

Dataset. We conduct experiments on the Deepfake Detection and Localization image dataset (DDL-I) [Miao *et al.*, 2025], which was originally released from the *IJCAI 2025 Challenge "The Deepfake Detection and Localization"*. The DDL-I dataset is a large-scale Deepfake benchmark comprising 1.2 million images with pixel-level annotations. The dataset includes both genuine and manipulated facial images, and each fake image is paired with pixel-level mask annotations that indicate the manipulated regions. The DDL-I dataset includes 61 representative Deepfake methods across four major forgery types: face swapping, face reenactment, full-face synthesis, and face editing. In addition, it encompasses both single-face and multi-face scenarios, simulating complex Deepfake content and contexts in real-world applications. Compared to existing datasets, DDL-I dataset provides superior diversity in forgery types, larger scale, and more complex scenarios, making it particularly suitable for both image-level Deepfake detection and fine-grained localization tasks. Following the original dataset splits, we use around 950k images for training, 240k for validation, and 220k for testing.

Evaluation Metrics. The performances of Deepfake detection and localization tasks are evaluated with different metrics.

Deepfake Detection. The detection task is evaluated using the standard AUC (Area Under the ROC Curve) metric for binary classification.

Deepfake Localization. The localization task is evaluated with two pixel-level metrics: F1-score and IoU (Intersection over Union).

The F1-score is calculated by:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \tag{8}$$

where TP, FP, and FN denote true positive, false positive, and false negative pixels, respectively.

The IoU measures the overlap between the predicted and ground-truth forgery regions:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}} = \frac{TP}{TP + FP + FN} \tag{9}$$

Table 1: Comparison results of Deepfake detection and localization on the DDL-I validation set. Bold indicates the best results.

Method	Detection	Localization	
	AUC	F1	IoU
HRNet-w18 [Wang *et al.*, 2020]	99.47	95.17	90.77
UperNet-Swin-T [Xiao *et al.*, 2018]	99.85	96.53	93.28
SegFormer-B5 [Xie *et al.*, 2021]	**99.92**	96.78	93.75
SAM [Kirillov *et al.*, 2023]	99.51	93.70	88.14
IML-ViT [Ma *et al.*, 2023]	-	94.10	90.20
Mesorch [Zhu *et al.*, 2025]	-	94.41	90.40
Ours	98.38	**97.90**	**95.88**

Implementation Details. The proposed solution is implemented using the MMSegmentation framework [Contributors, 2020]. We adopt the Swin-T backbone as a typical implementation of our framework. The backbone is initialized with pre-trained weights on ImageNet-1k [Russakovsky *et al.*, 2015]. Following [Cheng *et al.*, 2022], the number of layers L in the Transformer decoder is set to 9, and the number of learnable queries N is set to 100. For data preprocessing, each image is resized to a base scale of 2048×512 while preserving its original aspect ratio. For data augmentation, we first use random resizing with a scale factor of $[0.5, 2.0]$, and then use random cropping and padding to get training images with size 512×512. We also apply random flipping and typical photometric distortions, such as random brightness and contrast.

The proposed solution is trained end-to-end using annotations for both Deepfake detection and localization. Training is conducted on two NVIDIA Tesla A100 GPUs with a mini-batch size of 64. We use the AdamW optimizer with a learning rate of 0.0001, weight decay of 0.05, and a polynomial learning rate decay schedule for 30k iterations. Unless specialized otherwise, all compared methods adopt the same preprocessing and training settings.

4.2 Comparison to State-of-the-art Methods

We evaluate the proposed solution with Swin-T on the DDL-I validation set for both Deepfake detection and localization tasks, as presented in Table 1. We compare our approach with state-of-the-art semantic segmentation methods, including HRNet [Wang *et al.*, 2020], UperNet [Xiao *et al.*, 2018], SegFormer [Xie *et al.*, 2021], and SAM [Kirillov *et al.*, 2023]. To enable joint prediction of both tasks, we integrate an additional detection branch for these methods. We also include recent methods specifically designed for image manipulation localization, including IML-ViT [Ma *et al.*, 2023] and Mesorch [Zhu *et al.*, 2025]. For the Deepfake detection task, our method achieves a competitive AUC of 98.38%, comparable to existing methods. For Deepfake localization, our model obtains an F1 score of 97.90% and an IoU of 95.88%, surpassing state-of-the-art methods by a large margin. These results indicate that our approach can effectively achieve both accurate image-level Deepfake detection, as well as fine-grained localization of forgery regions.

Table 2: Ablation results of backbone selection and ensemble strategies on the DDL-I testing set.

Method	Detection	Localization	
	AUC	F1	IoU
Single backbone			
ResNet-50	88.32	72.76	66.58
Swin-T	93.42	77.78	71.79
Swin-S	94.24	76.18	70.79
Ensemble (Swin-T + Swin-S)			
Average	94.01	77.07	71.44
Re-weighted	94.95	-	-
Ours	94.95	77.78	71.79

4.3 Ablation Study

We conducted ablation studies to investigate the impact of backbone selection and ensemble strategies in our framework. Results on DDL-I testing set are shown in Table 2.

We first compare three single backbones: ResNet-50 [He *et al.*, 2016], Swin-T, and Swin-S [Liu *et al.*, 2021]. Among single backbones, Swin Transformer variants consistently outperform ResNet, indicating the effectiveness of shifted window attention in capturing global and local forgery cues. Notably, Swin-T achieves better localization performance (71.79% IoU), while Swin-S yields stronger detection accuracy (94.24% AUC). This is likely due to Swin-T, with a smaller receptive field, can provide finer-grained representation for Deepfake localization, while Swin-S, with a larger capacity, is more suitable for global Deepfake detection.

We further investigate two ensemble strategies based on Swin-T and Swin-S: vanilla averaging and re-weighted fusion. The simple averaging with equal ensemble weights performs even worse than using a single model. Then, we studied to use a re-weighted fusion strategy for Deepfake detection by assigning a higher weight ($w = 2.5$) to Swin-S model. This improves Deepfake detection AUC to 94.95%.

Our final solution (row 6) adopts this re-weighted strategy for detection and uses Swin-T for localization prediction. This combination achieves a promising overall performance: 94.95% AUC, 77.78% F1, and 71.79% IoU. By averaging these metrics, our solution achieved an overall score of 81.50% on the DDL-I testing set, securing a leading position in the *IJCAI 2025 Challenge "The Deepfake Detection and Localization"*. These results show the effectiveness of Swin Transformer and our tailored ensemble strategy in addressing key challenges in Deepfake detection and localization.

4.4 Visualization

We further visualize the results of our method on the Deepfake localization task. Here, we take the model with the Swin-T backbone as an example. The visualization results of the DDL-I validation set are illustrated in Figure 2. It covers three typical types of forgery scenarios: full-face forgeries (a), manipulations of specific facial components (b–c), and complex scenes involving multiple real and fake faces (d–e). The visualization results show that our method can accurately localize forgery regions under these cases, and the predicted

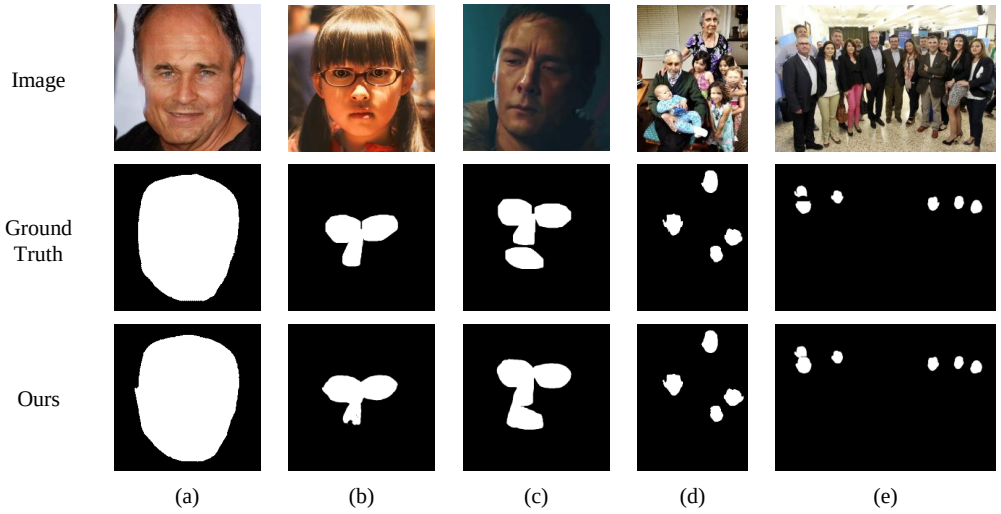

Figure 2: Visualization results of our method for Deepfake localization on DDL-I validation set. Our method consistently achieves accurate localization across diverse forgery cases, including full-face forgeries (a), manipulations of specific facial components (b–c), and complex scenes containing both real and fake faces (d–e).

masks are very close to the ground truth, indicating strong robustness and practical potential in real-world applications.

5 Conclusion

In this paper, we presented a unified framework for joint Deepfake detection and localization. Our solution utilizes a shared Swin Transformer backbone for feature extraction and integrates a masked-attention Transformer decoder for precise pixel-level Deepfake localization. The framework operates in an end-to-end manner without pre-processing steps, ensuring flexibility and simplicity. Extensive experiments on the DDL-I challenge benchmark demonstrate the effectiveness of our method, providing a practical solution for real-world application scenarios. In future work, we will explore the underlying relationship between Deepfake detection and localization tasks to enhance overall performance.

Acknowledgments

This work was supported by the Research Council of Finland (former Academy of Finland) Academy Professor project EmotionAI (grants 336116, 345122, 359854), HPC project FaceCanvas (grant number 364905), the University of Oulu & Research Council of Finland Profi 7 (grant 352788), and Infotech Oulu. As well, the authors wish to acknowledge CSC-IT Center for Science, Finland, for computational resources.

References

[Chai *et al.*, 2020] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–120. Springer, 2020.

[Cheng *et al.*, 2022] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022.

[Contributors, 2020] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[Dang *et al.*, 2020] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5781–5790, 2020.

[Haliassos *et al.*, 2021] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5039–5049, 2021.

[He *et al.*, 2016] 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 (CVPR)*, pages 770–778, 2016.

[Juefei-Xu *et al.*, 2022] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey, battleground, and horizon. *International Journal of Computer Vision (IJCV)*, 130(7):1678–1734, 2022.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

[Kong *et al.*, 2022] Chenqi Kong, Baoliang Chen, Haoliang Li, Shiqi Wang, Anderson Rocha, and Sam Kwong. Detect and locate: Exposing face manipulation by semantic-and noise-level telltales. *IEEE Transactions on Information Forensics and Security (TIFS)*, 17:1741–1756, 2022.

[Li *et al.*, 2021] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6458–6467, 2021.

[Li *et al.*, 2023] Chuqiao Li, Zhiwu Huang, Danda Pani Paudel, Yabin Wang, Mohamad Shahbazi, Xiaopeng Hong, and Luc Van Gool. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1339–1349, 2023.

[Lin *et al.*, 2024] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. Detecting multimedia generated by large ai models: A survey. *CoRR*, 2024.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[Liu *et al.*, 2022] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(11):7505–7517, 2022.

[Liu *et al.*, 2024] Haotian Liu, Chenhui Pan, Yang Liu, Guoying Zhao, and Xiaobai Li. Unified video and image representation for boosted video face forgery detection. In *European Conference on Artificial Intelligence (ECAI)*. IOS Press, 2024.

[Ma *et al.*, 2023] Xiaochen Ma, Bo Du, Zhuohang Jiang, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Benchmarking image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023.

[Miao *et al.*, 2023a] Changtao Miao, Qi Chu, Zhentao Tan, Zhenchao Jin, Wanyi Zhuang, Yue Wu, Bin Liu, Honggang Hu, and Nenghai Yu. Multi-spectral class center network for face manipulation detection and localization. *arXiv preprint arXiv:2305.10794*, 2023.

[Miao *et al.*, 2023b] Changtao Miao, Zichang Tan, Qi Chu, Huan Liu, Honggang Hu, and Nenghai Yu. F2trans: High-frequency fine-grained transformer for face forgery detection. *IEEE Transactions on Information Forensics and Security (TIFS)*, 18:1039–1051, 2023.

[Miao *et al.*, 2024] Changtao Miao, Qi Chu, Tao Gong, Zhentao Tan, Zhenchao Jin, Wanyi Zhuang, Man Luo, Honggang Hu, and Nenghai Yu. Mixture-of-noises enhanced forgery-aware predictor for multi-face manipulation detection and localization. *arXiv preprint arXiv:2408.02306*, 2024.

[Miao *et al.*, 2025] Changtao Miao, Yi Zhang, Weize Gao, Man Luo, Weiwei Feng, Zhiya Tan, Jianshu Li, Ajian Liu, Yunfeng Diao, Qi Chu, Tao Gong, Li Zhe, Weibin Yao, and Joey Tianyi Zhou. Ddl: A dataset for interpretable deepfake detection and localization in real-world scenarios. *arXiv preprint arXiv:2506.23292*, 2025.

[Nguyen *et al.*, 2024] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17395–17405, 2024.

[Ojha *et al.*, 2023] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24480–24489, 2023.

[Rossler *et al.*, 2019] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[She *et al.*, 2024] Huimin She, Yongjian Hu, Beibei Liu, Jicheng Li, and Chang-Tsun Li. Using graph neural networks to improve generalization capability of the models for deepfake detection. *IEEE Transactions on Information Forensics and Security (TIFS)*, 2024.

[Shiohara and Yamasaki, 2022] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18720–18729, 2022.

[Shuai *et al.*, 2023] Chao Shuai, Jieming Zhong, Shuang Wu, Feng Lin, Zhibo Wang, Zhongjie Ba, Zhenguang Liu, Lorenzo Cavallaro, and Kui Ren. Locate and verify: A two-stream network for improved deepfake detection. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 7131–7142, 2023.

[Tan *et al.*, 2023] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on

gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12105–12114, 2023.

[Wang *et al.*, 2020] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(10):3349–3364, 2020.

[Xiao *et al.*, 2018] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:12077–12090, 2021.

[Zhang *et al.*, 2024a] Yi Zhang, Weize Gao, Changtao Miao, Man Luo, Jianshu Li, Wenzhong Deng, Zhe Li, Bingyu Hu, Weibin Yao, Wenbo Zhou, et al. Inclusion 2024 global multimedia deepfake detection: Towards multi-dimensional facial forgery detection. *arXiv preprint arXiv:2412.20833*, 2024.

[Zhang *et al.*, 2024b] Yi Zhang, Changtao Miao, Man Luo, Jianshu Li, Wenzhong Deng, Weibin Yao, Zhe Li, Bingyu Hu, Weiwei Feng, Tao Gong, and Qi Chu. Mfms: Learning modality-fused and modality-specific features for deepfake detection and localization tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 11365–11369, 2024.

[Zhou *et al.*, 2023] Jizhe Zhou, Xiaochen Ma, Xia Du, Ahmed Y Alhammadi, and Wentao Feng. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22346–22356, 2023.

[Zhu *et al.*, 2025] Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 11022–11030, 2025.