Mesoscopic artifact learning-based Facial Forgery Detection and Localization for IJCAI 2025 DDL-I Challenge

Wei Li[1] , **Chujie Tang**[1] , **Jiang Yuan**[1,2] , **Zhonghua Zhao**[1] , **Bo Wang**[1,*]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China [2]YunQue AGI, Hangzhou, China

Abstract

This paper presents our submitted system to the IJCAI 2025 Workshop and Competition on "Deepfake Detection, Localization, and Interpretability" Challenge Track 1 (DDL-I). This track focuses on dual objectives: image-level forgery classification and pixel-level localization of manipulated regions. Given that tampering operations primarily focus on semantic editing of local facial objects, we formulate the deepfake detection task as a mesoscopic-level pattern recognition problem. The proposed system uses Transformer- and CNN-based encoder-decoders to extract macro and micro artifact characteristics respectively, and integrates them into mesoscopic artifact representations to achieve precise detection of object-level tampering. We evaluate our best system on the test set of the track, achieving a comprehensive score of 0.805.

1 Introduction

Recent rapid advances in deep learning, especially generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] and diffusion models (DMs)[Ho *et al.*, 2020], have elevated facial forgery techniques (e.g. face swapping, face reenactment, full-face synthesis, and face editing) to photorealistic levels of deception. Simultaneously, these deepfake technologies are being weaponized for malicious purposes—including fake news propagation, identity fraud, and political defamation—severely undermining global social trust system.

Both academia and industry have begun paying close attention to deepfake detection, releasing a series of related technologies and datasets. However, current solutions predominantly adopt binary classification frameworks, neglecting precise analysis of forged regions. This not only compromises the reliability of authenticity judgments but may also obscure underlying model biases. Consequently, the task of deepfake detection and localization (DDL) has been proposed and received widespread attention. This task requires models to not only identify whether a facial image has been manipulated but also segment the precise tampered regions. DDL presents greater challenges than binary classification, as it demands the model to understand the spatial distribution patterns of forgery - making it particularly valuable for scenarios such as forensic investigation and content moderation that require interpretability. Although some progress has been made, current DDL research remains fundamentally limited by insufficient datasets. As shown in Table 1, only two datasets [He *et al.*, 2021; Zhou *et al.*, 2021] released in the past five years provide manipulated region annotations. Furthermore, existing datasets are insufficient in terms of scale, forgery diversity and scenario diversity.

Fortunately, the recent "Deepfake Detection, Localization, and Interpretability" workshop and competition in IJCAI 2025 released a large-scale high-quality deepfake dataset [Miao *et al.*, 2025; Zhang *et al.*, 2024a; Miao *et al.*, 2024; Miao *et al.*, 2023; Zhang *et al.*, 2024b], which provides over 1.5 million forged face samples with pixel-level manipulation annotations, involving 61 forgery algorithms and covering both single-face and multi-face scenarios. This dataset has significantly advanced the development of DDL technology, effectively enhancing the interpretability of deepfake detection. We are very excited to participate in this competition and contribute to the establishment of a verifiable and traceable deepfake analysis system.

This paper presents our contributions to the IJCAI 2025 DDL-Challenge Track 1 for facial image detection and localization (termed DDL-I). Through systematic observation, we identify that most facial manipulations (e.g., identity swapping, gender transformation, and expression modification) aim to deceive the audience by editing object-level semantic components (e.g., mouth, eyes) of facial images. Given that such manipulations involve both macro-level semantic changes (e.g., identity, age) and micro-level consistency disruptions (e.g., color, texture), we propose analyzing forgery from a mesoscopic perspective, i.e., integrating macro and micro features into mesoscopic-level artifact representations to achieve precise detection of object-level semantic tampering. In constructing the DDL-I system, we primarily drew upon the open-source project Mesorch[1] [Zhu *et al.*, 2025] as our foundational framework. In summary, our system first encodes the macro and micro artifacts using the Transformer- and CNN-based backbone respectively; Then, in the decoding phase, the pixel-level forgery localization results are ob-

*Corresponding author: Bo Wang (wangbo@ia.ac.cn)

[1]https://github.com/scu-zjz/Mesorch

Dataset	Year	Tasks	Deepfake Methods	#Fake	Multi-Face
FaceForensics++ [Rossler *et al.*, 2019]	2019	Cla	4	4K videos	-
Celeb-DF [Li *et al.*, 2020]	2020	Cla	1	5K+ videos	-
DFDC [Dolhansky *et al.*, 2020]	2020	Cla	8	0.1M+ videos	-
ForgeryNet [He *et al.*, 2021]	2021	Cla/SL	15	1M+ images	-
FFIW [Zhou *et al.*, 2021]	2021	Cla/SL	3	10K videos	✓
OpenForensics [Le *et al.*, 2021]	2021	SL	3	0.1M image	✓
DeepFakeFace [Song *et al.*, 2023]	2023	Cla	3	90K images	-
DiffusionDeepfake [Bhattacharyya *et al.*, 2024]	2024	Cla	2	0.1M+ images	-
DF40 [Yan *et al.*, 2024]	2024	Cla	40	1M+ images	-
DDL-I (IJCAI 2025-DDL Challenge)	**2025**	**Cla/SL**	**61**	**1.5M+ images**	✓

Table 1: Comparison of existing deepfake datasets

tained by weighted fusion of multi-scale feature maps, and on this basis, the image-level authenticity predictions are obtained by global max pooling operation. In the training phase, we implement extensive data augmentation techniques and use cross-entropy loss to optimize all model parameters. During inferencing, we apply post-processing techniques such as TTA [Krizhevsky *et al.*, 2012] to further improve performance. Detailed implementations are described in Section 3 and Section 4. In the final assessment, our best-performing model achieves a comprehensive score of 0.805.

2 Related Work

2.1 Region-level forgery detection

Early approaches for tampering detection and localization in facial images primarily focused on region-level detection. Some methods leveraged object detection frameworks such as Faster R-CNN [Ren *et al.*, 2015] to quickly localize suspicious regions in face images. For instance, [Zhou *et al.*, 2018] proposed a two-stream Faster R-CNN network trained in an end-to-end manner to detect tampered regions in images. Meanwhile, some works adopted sequential modeling techniques such as LSTMs [Hochreiter and Schmidhuber, 1997] to better capture inter-region dependencies. [Bappy *et al.*, 2019] divided an image into patches, extracted resampling features, and applied LSTMs to model spatial dependencies across patches, thereby achieving automatic localization of tampered regions.

2.2 Pixel-level forgery detection

With the increasing demand for fine-grained forensic analysis, pixel-level tampering segmentation methods have attracted growing attention. These methods often employ fully convolutional networks (FCNs), attention mechanisms, or multi-scale feature fusion designs to produce binary masks that delineate manipulated regions. For example, [Dang *et al.*, 2020] introduced an attention-based refinement strategy to enhance feature representation for pixel-wise localization. Li et al [Li *et al.*, 2020] proposed the Face X-ray approach, which models blended boundaries to reveal the intrinsic structure of manipulated areas, effectively overcoming the limitations of traditional low-level forgery traces. [Liu *et al.*, 2022a] developed a dual-stream FCN that integrates multi-scale and multi-path information, significantly improving both accuracy and robustness through end-to-end training. [Wang *et al.*, 2023] further fused global features with local patch-level cues to capture subtle local distortions, enriching pixel-level localization capabilities. In recent years, multimodal and explainable detection methods have become research hotspots. [Shao *et al.*, 2024] proposed a hierarchical reasoning mechanism for multimodal image-text pairs, which localizes tampered regions in both images and text, adapting to new types of forgeries. [Xu *et al.*, 2024] introduced the FakeShield framework, an explainable image forgery detection and localization approach aided by large multimodal language models. Similarly, the SIDA framework proposed by [Huang *et al.*, 2024] achieves pixel-level localization of forged regions and provides natural language explanations by integrating visual and linguistic information, significantly enhancing the transparency and robustness of detection.

3 Method

3.1 Overall architecture

The framework of our system is shown in Figure 1, which includes four key modules: DCT-based frequency divider (D-FD), CNN-based microscopic artifact extractor (C-MiAE), Transformer-based macroscopic artifact extractor (T-MaAE), and adaptive weighting module (AWM). Specifically, D-FD first decomposes the input RGB image into high-frequency (HF) and low-frequency (LF) components, which are then concatenated with the original image along the channel dimension to generate high/low frequency-enhanced images. The HF/LF components are then fed into C-MiAE module and T-MaAE module respectively for encoding-decoding based artifact recognition, where C-MiAE is responsible for capturing fine-grained tampering traces from a micro perspective, while T-MaAE is responsible for revealing suspicious semantic patterns from a macro perspective. Both modules output feature maps at four different scales of the decoder. Subsequently, the AWM module performs weighted fusion of the multi-scale feature maps output by the two modules, generating a micro-forgery prediction mask and a macro-forgery prediction mask respectively. Note that this module takes the original RGB image and its HF/LF components as input, generating a series of normalized weight matrices, where each element in each matrix reflects the im-

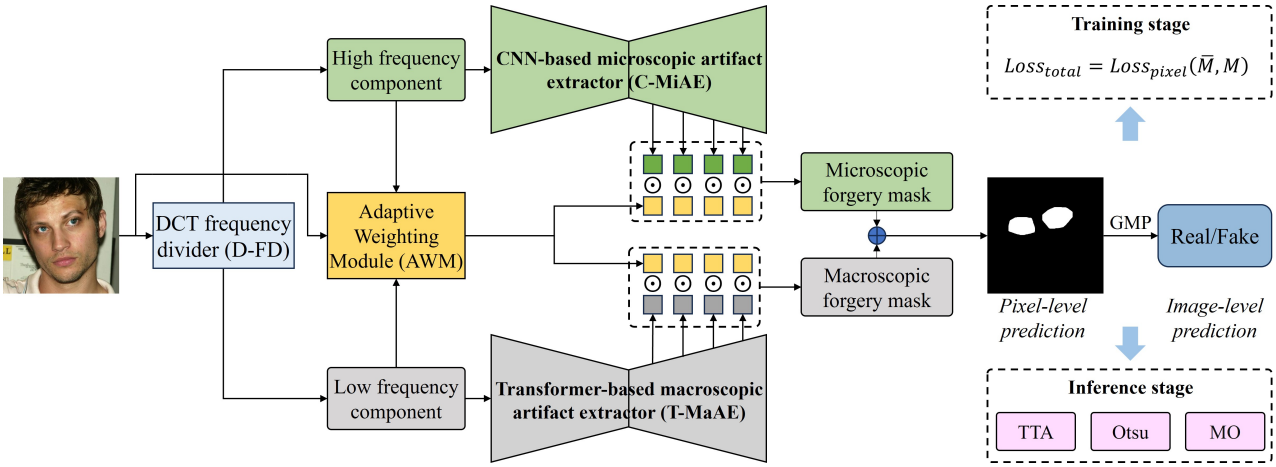

Figure 1: The framework of our system for IJCAI 2025 DDL-Challenge Track 1 (deepfake detection and localization, DDL-I).

portance of every pixel at the corresponding scale. Finally, the micro and macro prediction masks are fused to generate the final forgery mask. Then, based on the pixel-level forgery localization results, image-level authenticity prediction is obtained through a global max pooling (GMP) operation.

3.2 Model learning

The original Mesorch project employs a pixel-level forgery prediction head and an image-level forgery prediction head, and the total loss function of the model is the weighted sum of these two losses. As follows:

$$\begin{aligned} Loss_{\text{total}} = \alpha\, Loss_{\text{pixel}}(\bar{M}, M) \\ + (1-\alpha)\, Loss_{\text{image}}(H(\bar{M}), y) \end{aligned} \tag{1}$$

where $\bar{M}$ is the predicted pixel-level forgery mask, M is the groundtruth mask, H is the predicted image-level label and y is the groundtruth label. However, the dual loss (classification loss + mask loss) needs to balance classification confidence and localization accuracy, but there may be a conflict in optimization objectives between the two: The classification task focuses on global features while the localization task heavily relies on pixel-level supervision. Their weighted fusion not only requires manual weight adjustment (i.e. hyperparameter α) but also easily induces gradient competition, ultimately compromising model performance.

To address this issue, we replace the detection head with a global max pooling (GMP) layer, which is non-trainable. Consequently, the total loss is reduced to a single pixel-level mask-based entropy-cross loss:

$$Loss_{total} = Loss_{pixel}(\bar{M}, M). \tag{2}$$

$$\bar{y} = GMP(\bar{M}). \tag{3}$$

The single-loss strategy avoids suboptimal weight allocation and simplifies the training process. And according to experimental comparison, when using this single mask loss, the model performs better in minimizing missed detections of forged pixels and suppressing low-confidence artifacts. In addition, compared with the stacked structure of traditional convolutions and linear layers, GMP operation directly focuses on local salient forged features by extracting the maximum value from the spatial dimension. For example, forged regions often exhibit edge anomalies, inconsistent noise distributions, or texture discontinuities—such features typically manifest as local peak responses in activation maps. By preserving the maximum response of the feature map, GMP effectively suppresses irrelevant background interference and enhances the model's sensitivity to forgery traces.

4 Experiment

4.1 Data analysis

We have made a preliminary statistic on the size of the images in the competition training set. As shown in Figure 2, we found that all forged images can be mainly divided into the following three types:

- **Single small face image:** The image size is smaller than (384, 384), and there is only one face in the image. Such deepfake images account for approximately 12.6% of the entire training set, and most of their tampered regions are the entire face.
- **Single big face image:** The image size is larger than (384, 384), and there is only one face in the image. Such deepfake images account for approximately 32.6% of the training set, and most of their tampered regions are the eyes, nose, mouth, hair, etc.
- **Multi-face image:** The image has unequal length and width, and contains multiple faces. Such images account for approximately 54.8% of the training set, with most of their manipulations targeting the entire facial region.

4.2 Data augmentation

Dataset analysis reveals that single small face images account for a relatively small proportion in the training set, which may lead to insufficient learning of such forgery images. To alleviate this issue, we crop multiple-face images into individual single-face images, as shown in Figure 3. Specifically, we used the Deepface [Taigman *et al.*, 2014]

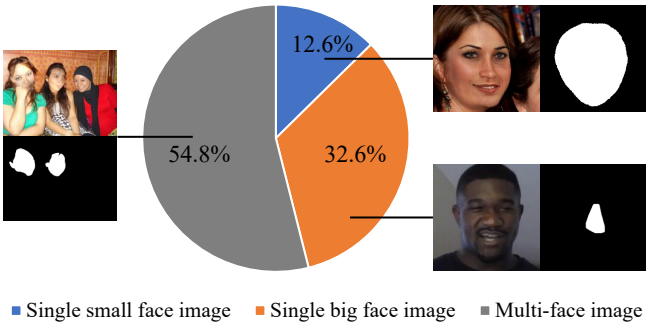

Figure 2: The statistics of the size distribution of fake faces in the training set.

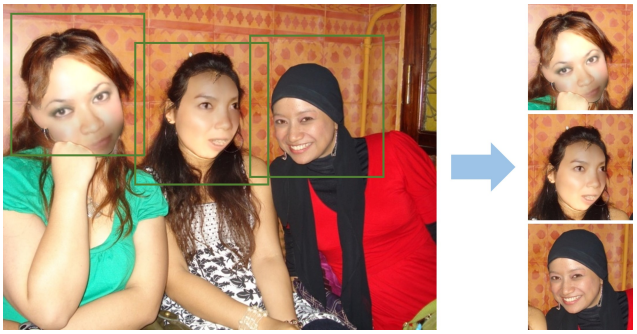

Figure 3: Expanding the single-face images by cropping multiple-face images.

tool to detect faces in the images, and then saved and obtained single small face images. In addition, we also use common image transformation methods for data augmentation, such as random cropping, random scaling, random shifting, etc.

4.3 Training details

Regarding model implementation, we chose ConvNext-base [Liu *et al.*, 2022b] as the backbone for C-MiAE and Segformer-B3 [Xie *et al.*, 2021] as the backbone for T-MaAE. We employ PyTorch (v2.6.0) as our training framework and utilize 8 NVIDIA RTX 4090 GPUs (24GB VRAM) for model training in this competition. The number of training epochs is set to 20, the batch size is set to 6 per GPU, and AdamW [Loshchilov and Hutter, 2017] is used as the optimizer. During training, all images are resized to 512×512. The initial learning rate is set to 1e-5, Cosine and Warmup are adopted as learning rate schedulers. The weight decay is set to 0.05, and the number of warm-up epochs is set to 2.

4.4 Inference details

To further enhance the model's performance on the three competition evaluation metrics (i.e. image-AUC, pixel-F1-score, and pixel-IoU), we implement three post-processing techniques as follows:

1) Morphological operations (MO) are a set of shape-based image processing techniques commonly used for post-processing of binary images or grayscale images. The main purpose is to change the shape, connectivity, or remove noise of objects through specific rules. Their core idea is to achieve fine-grained adjustments by sliding a predefined structural element (kernel) over the image and performing logical operations. We use the following three MO methods:

- **Erosion:** Erode the target edges with a structural element, shrink the white area, and eliminate isolated island-like noise.
- **Dilation:** Dilate the target edges with a structural element, expand the white area, and fill holes or broken areas.
- **Opening:** Erode first and then dilate to retain the main body and remove small noise.

2) Test-time augmentation (TTA) is a technique that enhances the robustness and accuracy of predictions through data augmentation during the model inference phase. Unlike using data augmentation only during training, TTA generates multiple augmented versions of the same input during testing and combines the prediction results of all versions as the final output. In this competition, our TTA ultimately uses three transformation methods: horizontal flip, vertical flip, and 90-degree rotation + horizontal flip. For the final mask, the maximum value of multiple prediction results is taken, and the forged confidence score is the average value.

3) Otsu algorithm is an automatic global threshold segmentation method based on grayscale histograms. Its core idea is to find the optimal segmentation threshold by maximizing the between-class variance of the foreground (object) and background, making it suitable for converting grayscale images into binary images. The binary segmentation of images can be quickly achieved through the Otsu algorithm.

4.5 Evaluation metrics

This DDL-I competition selects three evaluation metrics to test the performance of the participating systems. First, the Area Under the ROC Curve (AUC) is adopted to evaluate the detection performance. It measures the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across different thresholds. Second, the F1 Score and Intersection over Union (IoU) are adopted to evaluate the spatial localization performance. Specifically, the F1 Score is used to evaluate the balance between precision and recall, especially under imbalanced class distributions. It is the harmonic mean of precision and recall:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively. The IOU measures the overlap between the predicted and ground-truth manipulation regions:

$$\text{IoU} = \frac{|\text{Prediction} \cap \text{Ground Truth}|}{|\text{Prediction} \cup \text{Ground Truth}|}$$

These three metrics jointly assess both the classification accuracy and the spatial localization capability of the systems.

4.6 Ablation study

We conduct a series of ablation experiments on the competition-provided validation set to determine the optimal DDL-I system configuration, including backbone selection and loss function design. The details are as follows.

Local Feature Module	Global Feature Module	AUC	F1-Score	IoU	Avg.
ConvNeXt	Swin Transformer	0.9992	0.8016	0.7775	0.8594
	SegFormer	**0.9997**	**0.8019**	**0.7798**	**0.8605**
ResNet-50	Swin Transformer	0.9868	0.7776	0.7426	0.8357
	SegFormer	0.9985	0.7873	0.7539	0.8467

Table 2: Performance comparison of different backbone combinations on the validation set.

Model	Loss setting	AUC	F1-Score	IoU	Avg.
Mesorch	loc+cls	0.9845	0.7702	0.7322	0.8296
Ours	loc	**0.9997**	**0.8019**	**0.7798**	**0.8605**

Table 3: Performance comparison of different loss settings on the validation set.

4.6.1 Effect of different backbones

For the CNN-based microscopic artifact extractor (C-MiAE), we evaluate both ResNet-50 [He *et al.*, 2016] and ConvNeXt-Base [Liu *et al.*, 2022b] as potential backbones. For the Transformer-based macroscopic artifact extractor (T-MaAE), SegFormer-B3 [Xie *et al.*, 2021] and Swin Transformer-Base [Liu *et al.*, 2021] were selected for comparative testing. The performance comparison of different combinations of CNN backbone and Transformer backbone is shown in Table 2. It can be seen that the combination of ConvNeXt-Base and SegFormer-B3 excels in all evaluation metrics, outperforming other model combinations.

4.6.2 Effect of different loss settings

The original Mesorch project uses a weighted sum of pixel-level localization loss and image-level classification loss for model training, while we only utilize a pixel-level localization loss. Based on the optimal backbone combination mentioned above, we conduct a comparative analysis of these two schemes. The performance comparison of different combinations of CNN and Transformer backbone is shown in Table 3. It can be seen that our method outperforms the original Mesorch project in all metrics, showing the superiority of using a single localization loss. We further visualize the forgery localization effects of two schemes on the validation set. As shown in Figure 4, the masks generated by our method are clearer and closer to the groundtruth. Both quantitative and qualitative results demonstrate the superior choice of using a single pixel-level mask-based loss. This not only reduces the model complexity but also improves the detection robustness and localization accuracy.

4.6.3 Effect of different post-processing techniques

Furthermore, we demonstrate the contributions of different post-processing techniques to our system's performance. As shown in Table 4, all three post-processing techniques effectively enhance performance. Among them, TTA shows the most significant improvement, achieving a comprehensive score increase of 0.0084 compared to the base version. The combination of different techniques can further improve performance compared to using techniques alone, demonstrating their complementary nature.

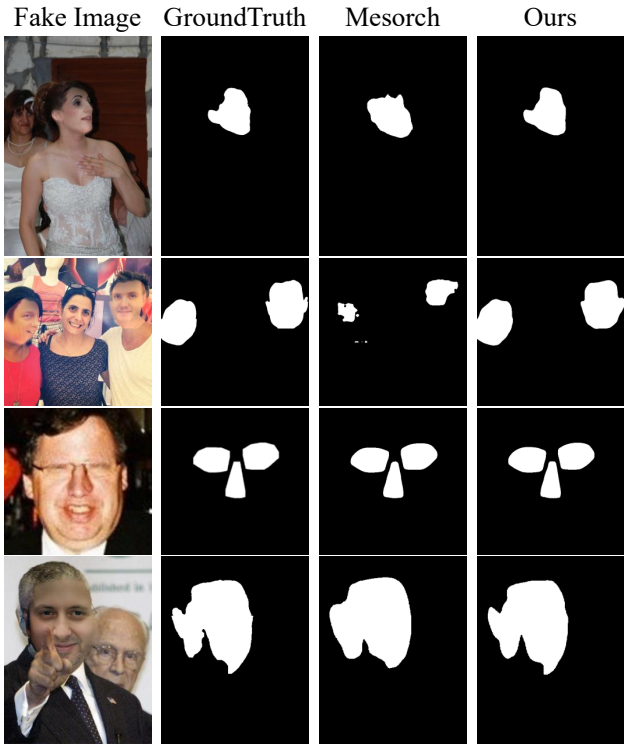

Figure 4: Qualitative analysis between our system and the original Mesorch on DDL-I task (no post-processing).

4.7 Comparison with state-of-the-art

Using the competition's dataset and evaluation metrics, we compare the constructed DDL-I system with existing advanced works (i.e., MVSS-Net [Chen *et al.*, 2021], Trufor [Guillaro *et al.*, 2023], and IML-ViT [Ma *et al.*, 2023]). All compared DDL-I methods meet two criteria: 1) Open-source code availability; 2) Capability for simultaneous detection and localization. As shown in Table 5, our system demonstrates significant advantages on the competition validation

Post-Proc.	AUC	F1-Score	IoU	Avg.
None (base)	0.9997	0.8019	0.7798	0.8605
+MO	0.9997	0.8037	0.7869	0.8634
+TTA	0.9998	0.8123	0.7947	0.8689
+Ostu	0.9997	0.8102	0.7913	0.8671
+MO+TTA	0.9998	0.8184	0.7978	0.8720
+MO+Ostu	0.9997	0.8121	0.7963	0.8685
+TTA+Ostu	0.9998	0.8211	0.7982	0.8730
+MO+TTA+Ostu	0.9998	0.8245	0.8038	0.8760

Table 4: Performance comparison of different post-processing techniques on the validation set.

Method	AUC	F1-Score	IoU	Avg.
MVSS-Net	0.8933	0.7748	0.7319	0.8000
Trufor	0.9547	0.7873	0.7529	0.8316
IML-Vit	0.9126	0.7920	0.7561	0.8202
Ours	**0.9997**	**0.8019**	**0.7798**	**0.8605**

Table 5: Performance comparison of different methods. All methods were trained and validated on the DDL-I competition dataset.

set, outperforming the second-best approach by 0.0403 in comprehensive performance (i.e., Avg. score). These results validate the effectiveness of mesoscopic-based facial artifact analysis framework.

Conclusion

In this work, we present our submitted system for the IJCAI 2025 DDL-Challenge Track 1 (Deepfake Detection and Localization). Our approach formulates facial object-level manipulation detection task as a mesoscopic pattern recognition problem. We accordingly use CNN- and Transformer-based encoder-decoders to integrate micro- and macro-artifact features into mesoscopic-level artifact representations, so as to achieve precise detection of object-level semantic tampering. Through meticulous parameter training and comprehensive post-processing during inference, our best system achieves a comprehensive score of 0.805 in the final official assessment.

Acknowledgments

This work is supported by the Natural Science Foundation of China (Grant No. 62192782).

References

[Bappy *et al.*, 2019] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE transactions on image processing*, 28(7):3286–3300, 2019.

[Chen *et al.*, 2021] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14185–14193, 2021.

[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*, pages 5781–5790, 2020.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Guillaro *et al.*, 2023] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20606–20615, 2023.

[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*, pages 770–778, 2016.

[He *et al.*, 2021] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4360–4369, 2021.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Huang *et al.*, 2024] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. *arXiv preprint arXiv:2412.04292*, 2024.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[Li *et al.*, 2020] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[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*, pages 10012–10022, 2021.

[Liu *et al.*, 2022a] Yumei Liu, Yong Zhang, and Weiran Liu. A novel face forgery detection method based on augmented dual-stream networks. In *Proceedings of the 2022 5th International Conference on Artificial Intelligence and Pattern Recognition*, pages 331–337, 2022.

[Liu *et al.*, 2022b] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[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.*, 2023] 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.*, 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.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[Shao *et al.*, 2024] Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. Detecting and grounding multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Taigman *et al.*, 2014] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[Wang *et al.*, 2023] Jun Wang, Benedetta Tondi, and Mauro Barni. Classification of synthetic facial attributes by means of hybrid classification/localization patch-based analysis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[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*, 34:12077–12090, 2021.

[Xu *et al.*, 2024] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*, 2024.

[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*, pages 11365–11369, 2024.

[Zhou *et al.*, 2018] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1053–1061, 2018.

[Zhou *et al.*, 2021] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5778–5788, 2021.

[Zhu *et al.*, 2025] Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Jizhe Zhou. Mesoscopic insights: Orchestrating multi-scale hybrid architecture for image manipulation localization. In *39th AAAI Conference on Artificial Intelligence*, pages 11022–11030, 2025.