Loupe: A Generalizable and Adaptive Framework for Image Forgery Detection

Yuchu Jiang[1,2*], **Jiaming Chu**[2,3*], **Jian Zhao**[2,5*], **Yuer Li**[2,6], **Xin Zhang**[2,4]
Xinru Wang[2,7], **Mingxing Yuan**[2,8], **Xu Yang**[1†], **Lei Jin**[3]
Chi Zhang[2], **Xuelong Li**[2†]
[1]Southeast University [2]TeleAI of China Telecom
[3]Beijing University of Posts and Telecommunications
[4]Lanzhou University [5]Northwestern Polytechnical University
[6]Dalian University of technology
[7]Beijing Institute of Technology
[8]Beijing University of Technology
{kamichanw, xuyang_palm}@seu.edu.cn, {chujiaming886, jinlei}@bupt.edu.cn,
3542558675@qq.com, xinzhang21@lzu.edu.cn, xr_vv123@163.com,
yuanmingxing25@126.com, {zhaoj90, zhangc120, xuelong_li}@chinatelecom.cn

Abstract

The proliferation of generative models has raised serious concerns about visual content forgery. Existing deepfake detection methods primarily target either image-level classification or pixel-wise localization. While some achieve high accuracy, they often suffer from limited generalization across manipulation types or rely on complex architectures. In this paper, we propose **Loupe**, a lightweight yet effective framework for joint deepfake detection and localization. Loupe integrates a patch-aware classifier and a segmentation module with conditional queries, allowing simultaneous global authenticity classification and fine-grained mask prediction. To enhance robustness against distribution shifts of test set, Loupe introduces a pseudo-label-guided test-time adaptation mechanism by leveraging patch-level predictions to supervise the segmentation head. Extensive experiments on the DDL dataset demonstrate that Loupe achieves state-of-the-art performance, securing the first place in the IJCAI 2025 Deepfake Detection and Localization Challenge with an overall score of 0.846. Our results validate the effectiveness of the proposed patch-level fusion and conditional query design in improving both classification accuracy and spatial localization under diverse forgery patterns. The code is available at *https://github.com/Kamichanw/Loupe*.

1 Introduction

Recent progress in generative AI [Croitoru *et al.*, 2023; Shuai *et al.*, 2024; Zhan *et al.*, 2023] has greatly enhanced the capability to produce high-quality, realistic images, thereby facilitating the creation of content that closely mimics the real world. However, these technological advances also raise significant concerns about potential malicious misuse, particularly in the fabrication of deceptive content aimed at misleading the public or altering historical narratives. In response to these risks, the computer vision community has been actively developing advanced deepfake detection methods. Contemporary methods [Lin *et al.*, 2024a; Yan *et al.*, 2023] primarily focus on evaluating the authenticity of the entire image (i.e., *real* or *forged*), while there is also an emerging subset dedicated to localizing tampered regions [Guo *et al.*, 2023; Li *et al.*, 2024].

Specifically, earlier approaches relied on visual networks (e.g., CNNs and ViTs) [Pei *et al.*, 2024; Guo *et al.*, 2023; Li *et al.*, 2024] or frequency-domain analysis [Pei *et al.*, 2024; Kwon *et al.*, 2022; Tan *et al.*, 2024] to extract features characteristic of images generated by GANs or diffusion models, aiming to detect or localize forgeries. However, these methods are typically architecturally complex and domain-specific, often exhibiting limited generalization to images produced by different generation techniques [Pei *et al.*, 2024; Lin *et al.*, 2024b]. On the other hand, recent studies have leveraged vision-language models (VLMs) [Huang *et al.*, 2024; Kang *et al.*, 2025], which simultaneously enable forgery detection and localization while offering interpretability, and have demonstrated strong performance. Nevertheless, the substantial computational resources required by VLMs constrain their practical deployment. Therefore, it is critical to develop a method that is computationally efficient, structurally simple, and capable of generalizing across various forgery techniques.

In this paper, we propose **Loupe**, a novel framework for image forgery detection and forged region localization, designed to simultaneously perform authenticity classification and precise localization of tampered regions. Loupe integrates an image encoder, a classifier, and a segmenter, jointly

*Equal contribution
†Corresponding author

modeling both authenticity verification and forgery localization tasks. To address the challenge of poor cross-domain generalization, we aim to introduce supervision signals at test time for dynamic adaptation. We note that image-level classification is generally less complex than pixel-wise segmentation, meaning the classification head often achieves better performance. Fortunately, with the advancement of large-scale visual pretraining, state-of-the-art vision backbones can be directly applied to dense prediction tasks without the need for complex segmentation networks [Bolya *et al.*, 2025; Tschannen *et al.*, 2025; Oquab *et al.*, 2023; Kerssies *et al.*, 2025]. Therefore, in our classifier, in addition to the traditional full-image prediction, we incorporate patch-wise predictions, resulting in a low-resolution mask prediction. This mask prediction can be used as a pseudo-label during testing, serving as a supervision signal to guide the segmenter. Additionally, patch-wise predictions can be combined with the traditional full-image predictions to yield the final result. This fusion strategy enhances the robustness and reliability of the image-level prediction.

We evaluate the effectiveness of Loupe on the DDL dataset [Organizers,]. On the validation set, the classification AUC reaches 0.946, while the segmentation IoU and F1 score attain 0.880 and 0.886, respectively. On the test set, the classification AUC reaches 0.963, and the segmentation IoU and F1 score are 0.756 and 0.819, respectively. Notably, the test set exhibits a mild distribution shift, containing some forgery techniques not present in the training set. Despite this, Loupe demonstrates robust performance, affirming the effectiveness of both the framework itself and the proposed test-time adaptation method.

2 Method

The overall architecture of Loupe is illustrated in Fig 1, comprising three primary components: the Image Encoder, the Classifier, and the Segmenter (comprising a Conditional Pixel Decoder and a Mask Decoder). The training process is conducted in two stages. Initially, the Image Encoder is frozen while the classification head is trained. In the second stage, the segmentation head is trained, still with the encoder frozen. The methodological details of each stage are presented in Sec 2.1 and Sec 2.2. Subsequently, Sec 2.3 describes how Loupe is employed for test-time adaptation.

2.1 Stage 1: Classification

In the first stage, we train the classification head to determine whether an input image is authentic or forged. Given an image $I \in \mathbb{R}^{3\times H\times W}$, where H and W represent the height and width of the image, respectively, the image is first processed by an image encoder to produce feature representations $\boldsymbol{F}_{16} \in \mathbb{R}^{H/16\times W/16\times D}$, where D denotes the output dimension of the image encoder, and we assume that the patch size of the image encoder is 16. The resulting feature map $\boldsymbol{F}$ is then passed to a patch-aware classifier.

The architecture of the patch-aware classifier is illustrated in Fig 2a. It begins with a pooling layer that aggregates global information from the entire image. This pooled representation is then passed through a multi-layer perceptron (MLP) to produce a global prediction. On the other hand, a separate MLP processes each image token individually to yield local predictions. Finally, a simple linear layer fuses both the global and local predictions to generate the final output $\hat{y}$.

In the image authenticity classification task, the number of forged patches is often substantially smaller than that of authentic patches. To mitigate the issue of class imbalance—where the majority class may dominate the learning process—we employ the poly focal loss $\mathcal{L}_{\text{patch}}$ [Leng *et al.*, 2022] as the supervision objective of patch prediction:

$$\mathcal{L}_{\text{patch}} = \frac{1}{N}\sum_{i=1}^{N}\left[-\alpha(1-p_i)^{\gamma}\log(p_i) + \epsilon(1-p_i)^{\gamma+1}\right]. \tag{1}$$

Here, $N = H/16 \times W/16$ is the number of total patches, p_i denotes the predicted probability for the forged class at patch i, α and γ are the focal loss coefficients, and ϵ is a scaling factor for the polynomial term. This formulation encourages the model to focus more on hard or underrepresented samples.

In addition, the global prediction is supervised using the standard binary cross-entropy loss $\mathcal{L}_{\text{global}}$. The final classification loss is defined as the sum of the patch-level and global losses:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{patch}} + \mathcal{L}_{\text{global}}. \tag{2}$$

2.2 Stage 2: Segmentation

In the second stage, we train the segmentation head to predict pixel-wise masks. Following *DetVit* [Li *et al.*, 2022], we apply a lightweight feature pyramid network (FPN) to the feature map $\boldsymbol{F}_{16}$ output by the image encoder, extracting multi-scale features at resolutions 1/4, 1/8, 1/16, and 1/32, resulting in $\{\boldsymbol{F}_4, \boldsymbol{F}_8, \boldsymbol{F}_{16}, \boldsymbol{F}_{32}\}$, where $\boldsymbol{F}_i \in \mathbb{R}^{D\times H_i\times W_i}$. For segmentation prediction, we adopt the *Mask2Former* [Cheng *et al.*, 2022] architecture. As the first step, to enhance the features, we employ a modified pixel decoder, referred to as the *Conditional Pixel Decoder*. In the i^{th} layer, the feature map $\boldsymbol{F}_i \in \{\boldsymbol{F}_4, \boldsymbol{F}_8, \boldsymbol{F}_{16}, \boldsymbol{F}_{32}\}$ is refined using multi-scale deformable attention (MSDA), outputting processed features $\tilde{\boldsymbol{F}}_i$. This process enables adaptive aggregation of information across multiple spatial resolutions while maintaining computational efficiency.

To support the pseudo-label-guided test-time adaptation introduced in Sec 2.3, the features output by MSDA are further processed through a cross-attention layer, where they interact with conditional queries, as illustrated in Fig 2b. These conditional queries not only guide the spatial aggregation but also incorporate high-level semantic information, allowing the subsequent mask decoder to produce semantically meaningful and spatially precise masks. In this process, the multi-scale features are transformed into a representation that is both resolution-consistent and semantically enriched.

The structure and training procedure of the mask decoder are consistent with those used in *Mask2Former*. Similar to patch classification in Sec 2.1, we supervise segmentation classification using poly focal loss instead of the standard binary cross-entropy. To further mitigate the issue of the model overly predicting authentic regions (e.g., false negatives), we adopt the Tversky loss $\mathcal{L}_{\text{tversky}}$ [Salehi *et al.*, 2017] as an auxiliary objective:

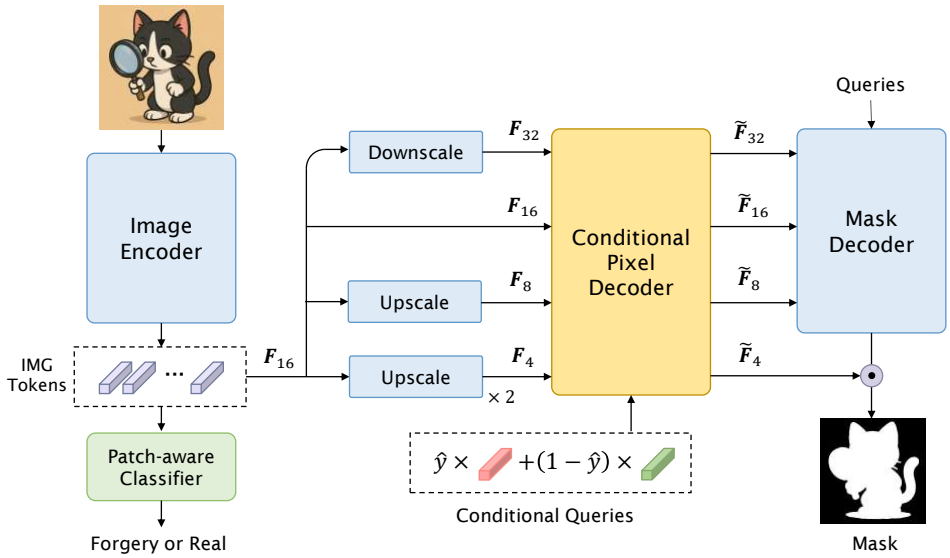
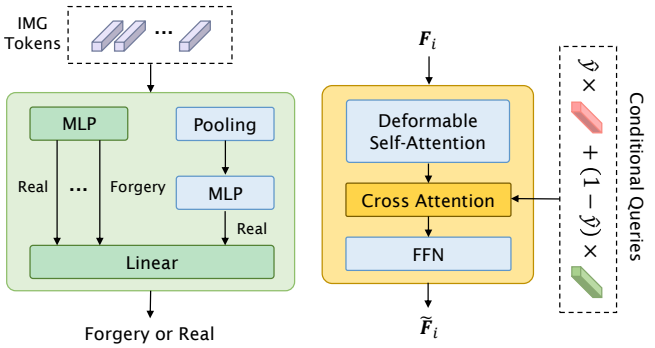


Figure 1: **The framework of Loupe.** Loupe consists of three main components: the image encoder, the classifier, and the segmenter. The image encoder is a vision backbone based on the ViT architecture. The classifier, which builds upon the traditional full-image prediction, extends it by adding patch-wise predictions. The segmenter follows the same meta-architecture as *Mask2Former*, with a key modification in the pixel decoder. The unchanged components are represented by blocks colored in light blue.



(a) Patch-aware classifier (b) Conditional pixel decoder layer

Figure 2: (a) **The detailed structure of the patch-aware classifier.** In addition to the conventional token prediction after pooling, Loupe also predicts the authenticity of each individual patch. (b) **The architecture of a conditional pixel decoder layer.** Loupe introduces a cross-attention layer between the deformable self-attention and the feed-forward network, which facilitates interaction with the conditional queries. For simplicity, residual connections and layer normalization are omitted.

$$\mathcal{L}_{\text{tversky}} = 1 - \frac{\text{TP}}{\text{TP} + \alpha \cdot \text{FP} + \beta \cdot \text{FN}}, \tag{3}$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. The coefficients α and β allow control over the penalty for FP and FN, enabling a trade-off between precision and recall. In our experiments, we set $\alpha = 0.3$ and $\beta = 0.7$ to prioritize recall and reduce missed detections of forged regions.

The overall loss function is formulated as:

$$\mathcal{L}_{\text{seg}} = \lambda_1 \mathcal{L}_{\text{mask}} + \lambda_2 \mathcal{L}_{\text{tversky}} + \lambda_3 \mathcal{L}_{\text{box}}, \tag{4}$$

where $\mathcal{L}_{\text{mask}}$ is the classification loss that uses poly focal loss to handle the class imbalance, giving more importance to the minority class (forged regions). $\mathcal{L}_{\text{box}}$ is the bounding box loss, where Hungarian matching is used to optimally assign predicted bounding boxes to the ground-truth forged regions, ensuring spatial consistency in the predicted forged areas.

2.3 Pseudo-Label-Guided Adaption

As discussed in Sec 1, previous forgery detection methods often lack generalization, making it difficult to apply them in real-world scenarios with out-of-distribution (OOD) data. Thus, investigating the application of trained models during testing is a critical challenge. To address this challenge, we propose a method for introducing supervision signals into the segmentation framework during testing.

To achieve this, during the training of Stage 2, we define two learnable embeddings to represent the two semantic categories: "authentic" and "forged". Based on the true labels, the corresponding embedding is selected to interact with the image features $\boldsymbol{F}_i$ in the conditional pixel decoder. During testing, we use the final output of the classifier as a pseudo-label, interpolating between the two semantic embeddings to provide additional conditions for the pixel decoder. The patch-level prediction results are treated as a low-resolution mask, which is subsequently passed into the mask decoder for supervision.

3 Experiments

3.1 Setup

Dataset and evaluation. We trained and evaluated Loupe on the DDL dataset[Miao *et al.*, 2025], which comprises both real/fake classification and spatial localization tasks. The

Table 1: **Leaderboard of the IJCAI 2025 Deepfake Detection and Localization Challenge**. The *overall* score is computed as the average of AUC, F1, and IoU.

Rank	AUC	F1	IoU	Overall
1 (ours)	0.963	0.756	0.819	**0.846**
2	-	-	-	0.8161
3	-	-	-	0.8151
4	-	-	-	0.815
5	-	-	-	0.815

Table 2: Ablation on patch prediction.

	AUC
Loupe (ours)	**0.946**
– patch prediction	0.920

dataset includes over 1.5 million images, covering 61 manipulation techniques, such as single-face and multi-face tampering scenarios. For evaluation, we used Area Under the ROC Curve (AUC) for detection, F1 Score, and Intersection over Union (IoU) for spatial localization (IoU is calculated exclusively for fake samples) as our metrics.

Implementation details. We use the *Perception Encoder* [Bolya *et al.*, 2025] as our image encoder. For the segmenter, most architectural parameters of the pixel decoder and mask decoder are kept consistent with those in *Mask2Former* [Cheng *et al.*, 2022], except that we set the number of learnable queries to 20. Each training stage runs for one epoch, using the AdamW optimizer. To adjust the learning rate, we adopt the warmup stable decay scheduler [Hu *et al.*, 2024], where the first 10% of the training steps are used for warmup and the final 10% for learning rate decay. More hyper-parameters are listed in Appendix A.

3.2 Results

Loupe secured first place in the IJCAI 2025 Deepfake Detection and Localization Challenge[Zhang *et al.*, 2024a; Zhang *et al.*, 2024b; Miao *et al.*, 2024; Miao *et al.*, 2023]. The top five entries on the leaderboard are shown in Table 1. Our method achieved an overall score that was 0.03 higher than the second-place entry, while the scores from second to fifth place differed by less than 0.001.

On the validation set, Loupe achieved a classification AUC of 0.947, and segmentation IoU and F1 scores of 0.880 and 0.886, respectively. Despite a slight distribution shift in the test set compared to the training and validation data, Loupe—particularly in classification AUC—remained largely unaffected, indicating the robustness of our approach.

3.3 Ablation Study

We conduct a series of ablation studies on the validation set of the DDL dataset to evaluate the effectiveness of our proposed method.

Patch-aware classifier. We validate the importance of patch-wise prediction by removing it. As shown in Table 2, the patch-wise prediction yields a significant improvement over the conventional global-only method, demonstrating the effectiveness of the local-global fusion strategy.

Table 3: Ablation on conditional queries of our modified pixel decoder and training objectives.

	F1	IoU
Loupe (ours)	**0.880**	**0.886**
– conditional queries	0.870	0.874

Conditional pixel decoder. Table 3 shows that Loupe benefits from our proposed conditional queries. By conditioning the image features with semantic embeddings before they are fed into the mask decoder, this approach not only enables test-time adaptation but also enhances the semantic alignment between the predicted masks and the underlying forgery types, leading to more accurate and context-aware localization.

4 Conclusion

In this work, we introduced **Loupe**, a unified and efficient framework for both deepfake detection and forged region localization. By integrating a patch-aware classifier with a conditional pixel decoder, Loupe enables robust global and local prediction with minimal architectural complexity. Furthermore, we propose a pseudo-label-guided test-time adaptation mechanism to improve generalization under distribution shifts. Extensive experiments on the DDL dataset demonstrate that Loupe achieves state-of-the-art performance, outperforming all competitors in the IJCAI 2025 Deepfake Detection and Localization Challenge.

References

[Bolya *et al.*, 2025] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

[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*, pages 1290–1299, 2022.

[Croitoru *et al.*, 2023] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[Guo *et al.*, 2023] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023.

[Hu *et al.*, 2024] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[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.

[Kang *et al.*, 2025] Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, et al. Legion: Learning to ground and explain for synthetic image detection. *arXiv preprint arXiv:2503.15264*, 2025.

[Kerssies *et al.*, 2025] Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your vit is secretly an image segmentation model. *arXiv preprint arXiv:2503.19108*, 2025.

[Kwon *et al.*, 2022] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022.

[Leng *et al.*, 2022] Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Xiaojie Shi, Shuyang Cheng, and Dragomir Anguelov. Polyloss: A polynomial expansion perspective of classification loss functions. *arXiv preprint arXiv:2204.12511*, 2022.

[Li *et al.*, 2022] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022.

[Li *et al.*, 2024] Qingming Li, Xiaohang Li, Li Zhou, and Xiaoran Yan. Adafl: Adaptive client selection and dynamic contribution evaluation for efficient federated learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649, 2024.

[Lin *et al.*, 2024a] 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. *arXiv preprint arXiv:2402.00045*, 2024.

[Lin *et al.*, 2024b] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16815–16825, June 2024.

[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.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Organizers,] IJCAI 2025 Organizers. Ijcai 2025 deepfake detection and localization challenge. https://deepfake-workshop-ijcai2025.github.io/main/index.html. Accessed: 2025-05-29.

[Pei *et al.*, 2024] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*, 2024.

[Salehi *et al.*, 2017] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.

[Shuai *et al.*, 2024] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024.

[Tan *et al.*, 2024] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5052–5060, 2024.

[Tschannen *et al.*, 2025] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[Yan *et al.*, 2023] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426*, 2023.

[Zhan *et al.*, 2023] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Ko-

rtylewski, Christian Theobalt, and Eric Xing. Multimodal image synthesis and editing: The generative ai era. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15098–15119, 2023.

[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.

Appendix

A Hyper-parameters

In this section, we list all hyper-parameters used during training and test-time adaption in Table 4, Table 5 and Appendix A. It is worth noting that the weighting factor α in $\mathcal{L}_{\text{patch}}$ (see Eq (1)) and $\mathcal{L}_{\text{mask}}$ (see Eq (4)) is determined empirically on a randomly selected subset of the dataset. For instance, in the training set of the DDL dataset, forged pixels account for approximately 20% of all pixels. Therefore, to address the class imbalance problem, we set $\alpha = 0.8$ in $\mathcal{L}_{\text{mask}}$, as shown in Table 5.

Table 4: The hyper parameters for stage 1: classification.

Param	Value
learning rate (lr)	5e-4
lr scheduler	warmup-stable-decay
warmup steps	10% total steps
decay steps	10% total steps
epoch	1
batch size	48
accumulative grad batches	8
optimizer	AdamW
weight decay	1e-3
grad clip	1.0
$\mathcal{L}_{\text{patch}}$	$\alpha = 0.85, \gamma = 2.0, \epsilon = 1.0$

Table 5: The hyper parameters for stage 2: segmentation. For parameters not mentioned, keep the same as stage 1 or *Mask2Former*.

Param	Value
learning rate (lr)	5e-4
batch size	40
accumulative grad batches	3
weight decay	5e-2
num_queries	20
$\mathcal{L}_{\text{mask}}$	$\alpha = 0.8, \gamma = 2.0, \epsilon = 1.0$
λ_1	5
λ_2	5
λ_3	2

Table 6: The hyper parameters for test-time adaption. For parameters not mentioned, keep the same as stage 2 or *Mask2Former*.

Param	Value
learning rate (lr)	1e-4
batch size	96
accumulative grad batches	1