A Lottery Ticket Hypothesis Approach with Sparse Fine-tuning and MAE for Image Forgery Detection and Localization

Jiaying Zhu*, Dong Li*, Xueyang Fu, Gege Shi, Jie Xiao, Aiping Liu, Zheng-Jun Zha†

University of Science and Technology of China
{zhujy53, dongli6, shigg, ustchbxj}@mail.ustc.edu.cn, {xyfu, aipingl, zhazj}@ustc.edu.cn

Abstract

The rise in sophisticated image forgery techniques, driven by advancements in image editing and generation, has posed new security challenges. Traditional methods, designed for specific tampering artifacts, struggle with out-of-distribution image forgery detection. In this paper, we propose a shift in paradigm, placing greater emphasis on the universal characteristics of authentic images, as opposed to solely focusing on specific forgery signals. We introduce an enhancement to the Masked Autoencoder (MAE), aptly termed the Forgery MAE (FMAE). This modification retains the inherent characteristics of natural images while integrating multi-source forgery information. Our implementation involves applying the lottery ticket hypothesis during pre-training to identify forgery-sensitive parameters, followed by their sparse fine-tuning to target the forgery detection and localization task. Concurrently, we develop a 'mixture of experts' noise extractor to compile multi-source forgery data. Our FMAE effectively extracts forgery features and shows strong resilience against unseen forgeries. Extensive experiments across multiple datasets confirm our method's superior accuracy and generalization capability over existing techniques.

Code — https://github.com/JulietChoo/FMAE

Introduction

The increasing accessibility and prevalence of image forgery, facilitated by advancements in image editing and generation techniques, have incited concern among the public due to the potential security and privacy threats they pose to individuals and society. For instance, advanced image forgery technology enables malicious users to manipulate images, fabricate fake news (Zhang et al. 2023, 2024), and even create false evidence in court (Lin et al. 2023). Simultaneously, the emergence of rapidly evolving technologies, such as diffusion models (Ho, Jain, and Abbeel 2020), has given rise to increasingly sophisticated methods of image forgery, which present obstacles to media security. Consequently, the development of effective and widely applicable image forgery detection and localization (IFDL) methods has become critical.

* Co-first authors contributed equally.

† Corresponding author.

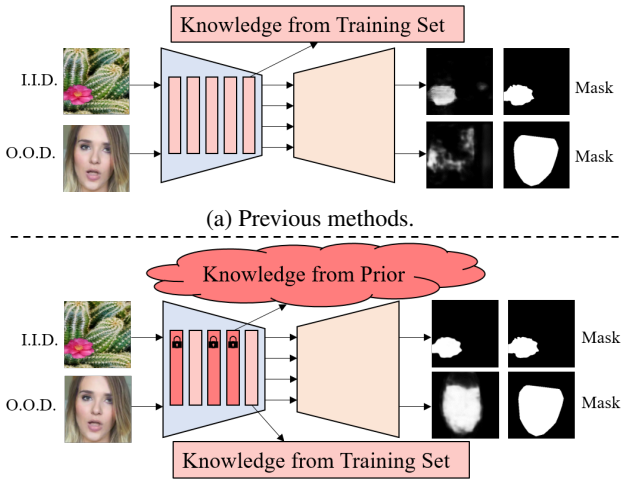


Figure 1: Comparison between previous methods and our proposed paradigm. Previous methods typically only use knowledge from the training data to extract tampering artifacts, resulting in poor out-of-distribution performance. Our paradigm retains natural image priors through sparse fine-tuning, allowing it to generalize to unseen forgeries.

Forensic techniques have made strides in recent years in combating forgery. These advancements include the exploitation of well-defined low-level features such as traces of JPEG compression, demosaicking, and interpolation (Kwon et al. 2022). Progress has also been made in specific areas of image tampering, such as splicing, with some studies achieving promising results (Yu et al. 2024). Furthermore, advancements have been made in the broader field of image tampering detection and localization. For example, RGB-N (Zhou et al. 2018) leverages noise features to model the inconsistencies between tampered and untouched regions while MVSS-Net (Chen et al. 2021c) utilizes both noise views and boundary artifacts to learn multi-view features. CAT-Net (Kwon et al. 2022) uses discrete cosine transform (DCT) coefficients, where compression artifacts remain, to localize image manipulation. However, as shown in Figure 1, the primary focus of most of these methods is the de-

sign of customized feature extraction networks for tampering artifacts, which compromises their out-of-distribution processing capability for forgery detection and localization. As a result, IFDL appears to be constantly trailing behind the evolution of forgery creation. While these methods are efficient at extracting outdated tampering artifacts, they struggle to keep pace with the rapidly evolving tampering methods.

In response, we propose an approach that emphasizes the general features of images rather than focusing exclusively on specific forgery clues. Unlike forged images, real-world images exhibit consistent naturalness (Nath, Gandhi, and Chouhan 2021). If a network can learn this consistent naturalness, it can detect previously unseen forgeries. To introduce such a prior in IFDL networks, we suggest using a model pre-trained on natural images. Specifically, we opt for the widely used Masked Autoencoder (MAE) (He et al. 2022) as the backbone. Demonstrating remarkable potential in various vision tasks, MAE can capture the relative position and semantics of images (Hu et al. 2023). Its ability to capture high-frequency details is crucial for image forgery detection and localization. Moreover, the pre-training paradigm of masked image modeling has proven superior to contrastive learning, another commonly used paradigm (Chen et al. 2023). Therefore, we aim to leverage the potent pre-trained weights of MAE to introduce the natural image prior.

However, integrating MAE into the downstream task of forgery detection and localization while preserving its natural image prior is a challenge. Either freezing the parameters of MAE makes it difficult for the encoder to extract forgery features, or unfreezing them all causes it to lose its natural prior. The Lottery Ticket Hypothesis (Frankle and Carbin 2019) offers a potential solution. The hypothesis, originally used for model pruning, seeks to find a subnetwork (or "winning tickets") that can replace the original dense network. Inspired by this, we posit that MAE contains a subset of parameters sensitive to forgery features, while another subset focuses on the consistent naturalness of images. After identifying the layers where forgery-sensitive parameters reside, introducing forgery-informed features to these layers during the fine-tuning stage can effectively enhance the network's forgery perception capability. This approach retains the generalization offered by the natural image prior and provides specificity in forgery perception.

Building on this, we introduce the forgery MAE with Lottery Ticket Sparse Fine-Tuning (LTSFT) for IFDL. The implementation involves identifying forgery-sensitive parameters during the pre-training stage using the lottery ticket hypothesis, followed by sparse fine-tuning of these parameters in downstream tasks related to forgery detection and localization. We also develop the 'mixture of experts' noise extractor to gather multi-source forgery information and feed it into the layers where forgery-sensitive parameters are located. Hence, our FMAE effectively extracts forgery features and demonstrates robust generalization ability against unseen forgeries, thanks to natural image priors. Our contributions are as follows:

- We present the Forgery Masked Autoencoder (FMAE), a new framework that harnesses the feature extraction capabilities of the Masked Autoencoder (MAE) and the natural image prior, to enhance the accuracy and generalizability of image forgery detection and localization.
- We introduce the Lottery Ticket Hypothesis and propose a strategy to identify forgery-sensitive parameters for sparse fine-tuning. This is the first application of this hypothesis in the context of IFDL. It is entirely orthogonal to existing methods and can help enhance their performance.
- We develop a 'mixture of experts' noise extractor to gather multi-source forgery information, thereby strengthening the forgery perception during the fine-tuning process.

Extensive experiments conducted on several representative benchmarks demonstrate that our method surpasses state-of-the-art methods, particularly in terms of generalization performance.

Related Work

Image Forgery Detection and Localization

The prior art in IFDL mainly relies on feature extraction and matching techniques, e.g. color filter array, photo-response non-uniformity noise, illumination, JPEG artifacts and so on. Despite the achievements, these methods often struggle with complex forgery techniques or when the forged region is well-blended into the background of the image.

In recent years, deep-learning techniques have attracted more attention (Peng et al. 2024; Ge, Fu, and Zha 2022; Ge et al. 2024). IFDL methods based on deep learning have achieved remarkable results (Ma et al. 2023; Zhu et al. 2024; Li et al. 2025). Many methods have been proposed to promote the progress and development of this field. For instance, MVSS-Net (Chen et al. 2021c) utilizes a dual-stream network with one stream dealing with noise distribution and the other designed edge-supervised branch to capture fine-grained boundary detail in a shallow to deep manner. PSCC-Net (Liu et al. 2022) uses a two-path (top-down and bottom-up route) methodology to analyze the image. TruFor (Guillaro et al. 2023) relies on the extraction of both high-level and low-level traces through a transformer-based fusion architecture that combines the RGB image and a learned noise-sensitive fingerprint. In this work, thanks to the utilization of natural image prior, we emphasize the universal features of authentic images instead of focusing only on specific forgery cues, thus improving generalizability.

Masked Image Modeling

This direction attempts to learn representations from images corrupted by masking. The earliest work DAE (Vincent et al. 2010) presents masking as a noise type and stacks layers of denoising autoencoders to denoise corrupted inputs. Most recently, inspired by the impressive performance of applying transformer architecture in natural language processing (NLP), transformer-based methods are developed. iGPT (Chen et al. 2020) adopts a sequence Transformer operating on sequences of pixels to auto-regressively predict masked pixels. BEiT (Bao et al. 2021) proposes to recover

the original discrete visual tokens based on the corrupted image patches. MAE (He et al. 2022) develops an asymmetric encoder-decoder architecture, with an encoder that operates only on the visible subset of patches, along with a lightweight decoder that reconstructs the original image from the latent representation and mask tokens. In this work, we leverage masked image modeling pre-training and the lottery ticket hypothesis to pick forgery-sensitive parameters and perform sparse fine-tuning, which allows our approach to combine both the feature extraction capabilities of MAE and the natural image prior.

Lottery Ticket Hypothesis

The lottery ticket hypothesis (LTH) (Frankle and Carbin 2019) is derived from neural network pruning technology, whose purpose is to prune away the excess capacity of a deep network. The LTH states: A randomly initialized, dense neural network contains a subnetwork that is initialized such that—when trained in isolation—it can match the test accuracy of the original network after training for at most the same number of iterations. LTH has been widely explored in various fields, such as natural language processing (Liang et al. 2023; Panda et al. 2024), generative adversarial networks (Chen et al. 2021d), image classification (Chen et al. 2021a), graph neural networks (Chen et al. 2021b), and reinforcement learning (Yu et al. 2019). However, most of them are still used for weight magnitude pruning. One relevant work (Ansell et al. 2022) to ours is from the natural language processing (NLP) field: the authors found that a simple variant of the LTH can accomplish task-specific and language-specific fine-tuning, which can prevent catastrophic forgetting. In contrast, we transplant the LTH-based fine-tuning to the image forensics task in visual domain. We design a forgery-aware sparse fine-tuning method based on masked image modeling and maintain the image prior of MAE to improve the accuracy and generalization for IFDL.

Methodology

Overview

Figure 2 provides an overview of our method. We first identify forgery-sensitive parameters in the MAE encoder through mask image modeling pre-training. Next, we fine-tune these parameters and introduce multi-source forgery features into specific layers. These steps constitute our lottery ticket sparse fine-tuning (LTSFT) and multi-source forgery awareness fine-tuning (MSFFT). Drawing on the Lottery Ticket Hypothesis (LTH) (Frankle and Carbin 2019), LTSFT identifies forgery-relevant parameters, freezes others, and fine-tunes only the relevant ones. MSFFT enhances LTSFT by injecting multi-source noise into selected layers, amplifying the network's sensitivity to forgery.

Formally, let I^F be forged images and I^A authentic images. Using MAE parameters trained on ImageNet as a base, we conduct mask image modeling on I^F and I^A to obtain parameter sets θ^F and θ^A. LTSFT then derives forgery-sensitive parameters θ^P, which are sparsely fine-tuned for the downstream IFDL task. Multi-source forgery information, extracted via a 'mixture of experts' noise extractor, is introduced into the network through these parameters, enhancing fine-tuning. The fine-tuned encoder outputs discriminative features, which are fed into a segmentation decoder to produce a forgery localization map. Finally, we employ ConvGeM from MVSS-Net (Dong et al. 2022) for image forgery detection.

Lottery Ticket Sparse Fine-Tuning

Finding Winning Tickets. We claim that the original parameters of the MAE encoder have certain natural image priors. Therefore, we perform pre-training on this basis to select forgery-sensitive parameters (winning tickets). This screening process can be divided into two phases.

In the first phase, let θ^H be the parameters of the MAE encoder trained by He et al. (He et al. 2022) on ImageNet, and θ^H is fully trained on the forged image dataset I^F for masked image modeling until convergence to yield θ^F. Then we rank the parameters according to the maximum absolute difference $\left|\theta_i^F - \theta_i^H\right|$, and select the top K to form the parameter set θ_{top}^F.

In the second phase, we still initialize with the original parameters θ^H and train the MAE until convergence using the authentic image dataset I^A corresponding to I^F to obtain θ^A. In the same way, we select the top K parameters in θ^A to form the parameter set θ_{top}^A based on the maximum absolute difference $\left|\theta_i^A - \theta_i^H\right|$. Then, forge-sensitive parameters θ^P which are also known as winning tickets can be picked out based on θ_{top}^F and θ_{top}^A. In our case, we use the difference set approach, which is computed as follows:

$$\theta^P = \theta_{top}^F - \theta_{top}^F \cap \theta_{top}^A, \tag{1}$$

where $-$ denotes the operation of finding the difference set, and $\cap$ denotes finding the intersection set.

Sparse Fine-Tuning. Inspired by ComposableSFT (Ansell et al. 2022), after resetting the parameters of the MAE encoder to the original parameters θ^H, we transfer the encoder to our image forgery detection and localization task. The forgery decoder is fully trained, while the encoder is sparsely fine-tuned. Specifically, only the selected parameters θ^P in the encoder are trainable, while other parameters are frozen. We generate a gradient mask m_p based on θ^P:

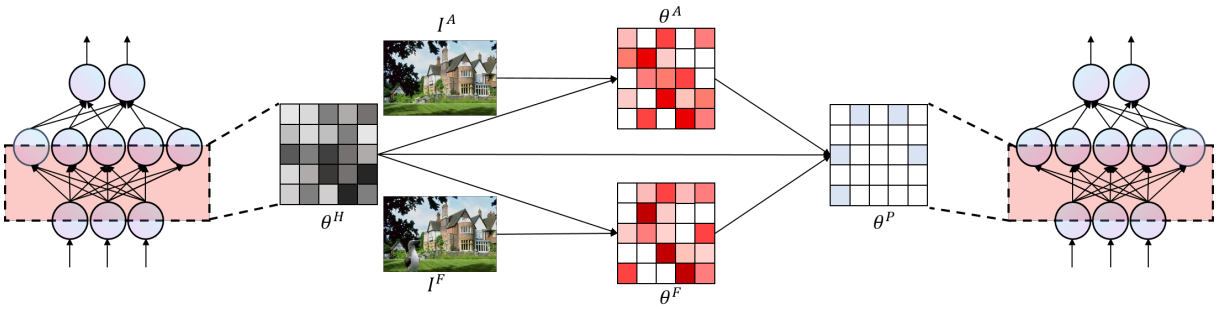
$$m_p\left(\theta_i\right) = \begin{cases} 0 & \theta_i \notin \theta^P \\ 1 & \theta_i \in \theta^P, \end{cases} \tag{2}$$

where θ_i is a parameter of the MAE encoder, and m_p is a binary mask with the same shape as the encoder parameter matrix. Based on this, we can calculate the masked gradient g_m, which is written as

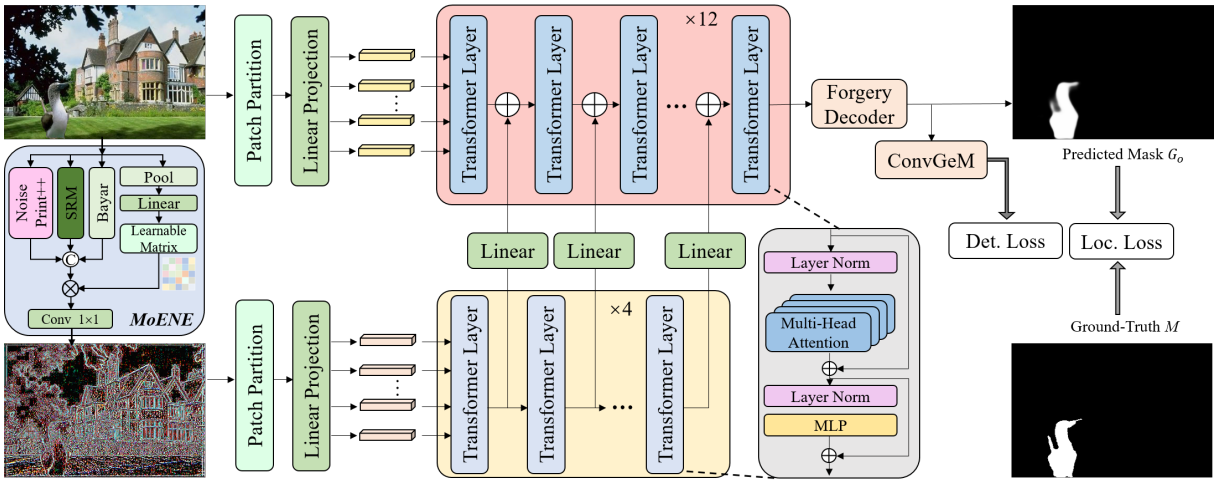
$$g_m = m_p \odot \nabla_\theta \mathcal{L}\left(F\left(\cdot, \theta, \delta\right)\right), \tag{3}$$

where $\odot$ is the Hadamard product, $\mathcal{L}$ is the loss function for IFDL, F is the overall network, θ are the parameters of the encoder, ∇_θ is the gradient of θ, and δ are parameters of the decoder which are optimized normally. Then, we pass g_m to the optimizer at each step to achieve sparse fine-tuning of the encoder.

It is worth noting that the proposed LTSFT is inspired by the Lottery Ticket Hypothesis (LTH) (Frankle and Carbin



(a) First stage of our method. This stage is used to find the parameters sensitive to image forgery in the MAE encoder through mask image modeling pre-training with the lottery ticket hypothesis.



(b) Second stage of our method. This stage is to sparsely fine-tune the found forgery-sensitive parameters and introduce multi-source forgery features into the layers where these parameters are located.

Figure 2: An overview of our Forgery MAE.

2019), but it makes unique changes based on the core idea of LTH to migrate it from pruning to sparse fine-tuning. After training a pre-trained model, we select the subset of parameters with the highest variation. Then, we rewind the model to its pre-training initialization (unlike the original LTH algorithm, we do not set any parameter values to zero). LTSFT achieves sparse fine-tuning in the form of difference vectors relative to pre-trained models by retraining only the selected subset of parameters.

Multi-Source Forgery Awareness Fine-Tuning

Previous research has found that exploring semantic information from images performs well for forgery detection within a distribution, but has poor Out of Distribution (OOD) performance (Zhou et al. 2020). In addition, introducing noise to learn content-agnostic information yields powerful forgery detection capabilities (Bayar and Stamm 2018; Hu et al. 2020; Fridrich and Kodovsky 2012). Based on the above, we speculate that relying solely on content-related information is insufficient to accurately detect forgery. However, previous methods usually rely on a single operator and strong supervision at the pixel level, which could limit the generalization ability of the model. To alleviate this problem, we propose multi-source forgery awareness fine-tuning (MSFFT), which is performed together with LTSFT.

MoENE. We design the 'Mixture of Experts' Noise Extractor (MoENE) to perform a unified co-exploration towards multi-source noise forgery information. We adopt three noise extraction algorithms as three experts, including SRM filter (Zhou et al. 2018; Fridrich and Kodovsky 2012), Bayar convolution (Bayar and Stamm 2018), and NoisePrint++ (Guillaro et al. 2023). These three algorithms are chosen because of their wide application (Zhou et al. 2018; Wu, AbdAlmageed, and Natarajan 2019; Dong et al. 2022; Li et al. 2023; Guillaro et al. 2023). SRM filter and Bayar convolution are two operators, and NoisePrint++ is a noise extractor based on contrastive learning. MoENE can improve performance while helping to alleviate overfitting by integrating the information extracted by the experts.

Formally, given an input image $X \in \mathbb{R}^{H \times W \times C}$, we first apply the channel-wise average to generate C-dimensional channel descriptor $T_c \in \mathbb{R}^C$:

$$T_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X\left(i, j\right), \tag{4}$$

where $X(i,j)$ denotes the (y,x) position of the X. Then we need two learnable weight matrices as coefficient vectors for each expert. We use a linear layer to get the first matrix $W_1 \in \mathbb{R}^{E\times C}$. Here, E is the dimension of the weight matrix. The process is written as

$$W_1 = \psi\left(\mathrm{Pool}\left(\mathcal{C}_3\left(X\right)\right)\right), \tag{5}$$

where ψ is the linear layer, Pool is the pooling operation and $\mathcal{C}_3$ is 3×3 convolution, which is used to reduce the computational complexity. The second matrix $W_2 \in \mathbb{R}^{OC\times E}$ is the learnable parameter matrix, where O is the number of experts, and $O = 3$ in practice. Then, we utilize these two weight matrices to adaptively process the outputs G_s, G_b, and G_n of the SRM filter, Bayar convolution, and Noiseprint++:

$$W_e = W_2\sigma\left(W_1 T_c\right), \tag{6}$$

$$G_f = \mathcal{C}_1\left(\mathrm{Cat}\left(G_s, G_b, G_n\right)W_e\right), \tag{7}$$

where $G_f \in \mathbb{R}^{H\times W\times C}$ denotes the obtained multi-source forgery feature, $W_e \in \mathbb{R}^{OC\times 1}$ denotes the coefficient matrix of the experts, and σ denotes a ReLU function. $\mathrm{Cat}()$ is the channelwise concatenation, which outputs a feature of shape $H\times W\times OC$, and $\mathcal{C}_1$ is 1×1 convolution. With this design, MoENE is able to adaptively extract forgery-informed noise from multiple sources. It is worth noting that MoENE does not attach an external gating network, which is different from the conventional 'mixture of experts' (Jacobs et al. 1991; Ren et al. 2018). Instead, we let the self-attention (Hu, Shen, and Sun 2018) become a switcher of different experts to adaptively select the importance of diverse representations depending on the inputs.

MSFFT. After obtaining the multi-source forgery information $G_f \in \mathbb{R}^{H\times W\times C}$, we add it to the encoder to perform multi-source forgery awareness fine-tuning (MSFFT). MSFFT is based on LTSFT. Since our LTSFT only fine-tunes winning ticket parameters, directly inputting the forged features into the layers with the most winning ticket parameters facilitates network optimization. Therefore, we first select the layers that need to add forgery information based on the winning ticket parameters θ^P in the Sec.. By counting the number of layers where θ^P are located, we select the N layers with the largest number of θ^P. Let the features of the N layers of the encoder be $G_{win} = \left\{G^1_{win}, \cdots, G^N_{win}\right\}$. In practice, layers 8, 9, 10 and 11 of the encoder are selected, so $G_{win} = \left\{G^1_{win}, G^2_{win}, G^3_{win}, G^4_{win}\right\} = \left\{G^8_r, G^9_r, G^{10}_r, G^{11}_r\right\}$, where G^i_r denotes the output of the i-th layer of the encoder. Meanwhile, we use N transformer blocks $\mathrm{Trans} = \left\{\mathrm{Trans}^1, \cdots, \mathrm{Trans}^{\mathrm{N}}\right\}$ to process noise features G_f. The process can be written as:

$$G^k_f = \mathrm{Trans}^k\left(G^{k-1}_f\right), \tag{8}$$

$$G^{k+1}_{win} = \mathrm{E}^i\left(G^k_{win}\right) + \varphi^k\left(G^k_f\right), \tag{9}$$

where $k \in \{1, 2, \cdots, N\}$, in practice $N = 4$. G^k_f is the processed noise feature, where $G^0_f = G_f$ and E^i is the corresponding layer in the MAE encoder. φ^k denotes the linear layer with both weight and bias initialized to zeros, which is to alleviate the impact on the natural image prior. Besides, φ^k converts G^k_f to the same shape as G^k_{win}.

Forgery decoder. During the fine-tuning process, we feed the feature output from the MAE encoder to the forgery decoder to predict the localization map. The design of the forgery decoder is not the focus of this paper, thus we utilize the commonly used segmentation decoder (Zheng et al. 2021). It outputs the predicted localization map G_o through progressive upsampling. As for forgery detection, we employ the ConvGeM proposed by MVSS-Net++ (Dong et al. 2022) to obtain the final binary prediction $\hat{y}$.

Loss Functions

Following MAE (He et al. 2022), in the pre-training stage, we use the Mean Squared Error between the reconstructed and original images in the pixel space on masked patches as the loss:

$$\mathbf{L}_{pre} = ||\mathrm{D}(P) - P_{un}||_2, \tag{10}$$

where P represents the mask tokens. Each mask token is a learned vector that indicates the presence of a patch to be predicted. P_{un} is the patches in the original image. $\mathrm{D}()$ denotes the MAE decoder. In the fine-tuning stage, the network aims to detect and localize forgeries.

The loss function of fine-tuning is written as:

$$\mathbf{L}_f = \lambda_1\mathcal{L}_1\left(y, \hat{y}\right) + \lambda_2\mathcal{L}_2\left(M, G_o\right) + \lambda_3\mathcal{L}_3\left(M, G_o\right), \tag{11}$$

where $\mathcal{L}_1$ and $\mathcal{L}_3$ denote the BCE loss, $\mathcal{L}_2$ denote the Dice loss, y is a label that represents the authenticity of the image, M is the ground-truth mask, and λ_1, λ_2 and λ_3 are the hyperparameters.

Experiment

Experimental Setup

Datasets. Following DiffForensics (Yu et al. 2024), we use both Casiav2.0 (Dong, Wang, and Tan 2013) and FantasticReality (Kniaz, Knyaz, and Remondino 2019) as training sets to select and fine-tune of forgery-sensitive parameters. The test sets include Casiav1.0 (Dong, Wang, and Tan 2013), Columbia (Hsu and Chang 2006), NIST16 (Guan et al. 2019), IMD2020 (Novozamsky, Mahdian, and Saic 2020), DSO-1 (De Carvalho et al. 2013), Korus (Korus and Huang 2016), AutoSplice (Jia et al. 2023), and OpenForensic (Le et al. 2021), of which the first six datasets were tampered with using traditional image editing tools, while the latter two datasets were tampered with using Deep Generative Models (DGMs).

Implementation Details. In this paper, the default framework is to unfreeze seventy percent of the parameters of the MAE encoder (ViT-B/16) for sparse fine-tuning while adding multi-source forgery information to its 8, 9, 10, and 11 layers. The networks are trained for 200 epochs with a batch size equal to 8. The input size is 512×512. Our model is trained with NVIDIA 3090. An AdamW optimizer with layer-wise learning rate decay is used. For the pre-training of the masked image modeling in the stage of picking the

Methods	Editing												DGM				Average	
	Casiav1.0		Columbia		NIST16		IMD2020		DSO-1		Korus		AutoSplice		OpenForensics			
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
ManTraNet*	.136	.612	.357	.767	.160	.741	.180	.785	.089	.687	.104	.681	.192	.622	.043	.678	.158	.697
SPAN	.088	.533	.213	.597	.116	.648	.108	.671	.059	.564	.070	.575	.047	.572	.014	.682	.089	.605
MVSS-Net*	.451	.845	.665	.818	.292	.791	.264	.817	.271	.732	.095	.641	.333	.839	.056	.702	.303	.773
CAT-Net	.394	.788	.854	.826	.336	.780	.295	.823	.135	.713	.149	.672	.185	.796	.003	.552	.294	.744
PSCCNet	.355	.738	.672	.881	.238	.740	.295	.800	.318	.721	.156	.623	.150	.784	.065	.610	.281	.737
HiFi-Net	.092	.642	.382	.608	.172	.685	.178	.675	.304	.700	.088	.607	.613	.831	.149	.676	.247	.678
DiffForensics	.517	.868	**.912**	.931	.415	.828	**.511**	**.911**	**.485**	.874	.257	.721	.507	.940	.122	.820	.466	.862
Ours	**.612**	**.876**	.881	**.935**	**.418**	**.846**	.491	.894	.469	**.896**	**.276**	**.736**	**.639**	**.950**	**.163**	**.844**	**.493**	**.872**

Table 1: Pixel-level F1 and AUC performance of image forgery localization. Method with * uses the pre-training model of the original paper.

forgery-sensitive parameters, we follow exactly the experimental setup of (He et al. 2022) for training. For the fine-tuning stages, we set $(\lambda_1, \lambda_2, \lambda_3) = (0.5, 0.35, 0.15)$ to balance detection loss and localization loss and set the initial learning rate as 5×10^{-5}.

Toy Experiment

To illustrate the strong forgery discrimination of our method, a toy experiment is performed: two thousand random samples are sampled in the test set to validate our method using only Casiav2.0 training. It is worth noting that in the toy experiments, we train the network using only authentic-forged labels, i.e., no localization loss is used. With such weak supervision, we visualize the obtained features as shown in Figure 3. It shows that our method pays attention to the forged region even under very weak supervision. We also perform t-SNE dimensionality reduction analysis on the authentic and forged image features in the test set. Compared with the fully unfrozen MAE, the model with sparse fine-tuning can better separate the real and fake images in the test set. As shown in Figure 4a, fine-tuning has been applied to forged images, involving training all parameters. In Figure 4b, only the winning ticket parameters are fine-tuned, demonstrating the efficiency of LTSFT, as it outperforms 4a. Compared with the fully unfrozen MAE, the model with sparse fine-tuning can better separate the forged images from the real images in the test set. This demonstrates the effectiveness of natural image priors for IFDL.

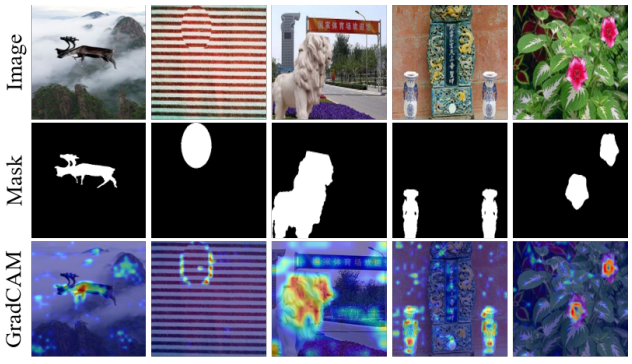
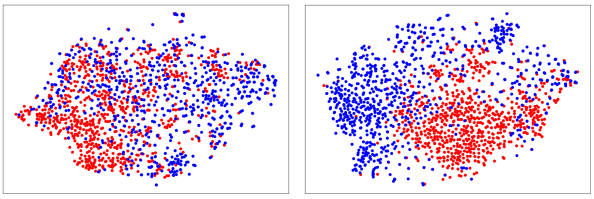


Figure 3: Feature visualization under weak supervision.



(a) Fully unfrozen MAE (b) Sparse fine-tuning MAE

Figure 4: Comparison of t-SNE analysis between fully unfrozen and sparse fine-tuning. The red dots represent authentic images, while the blue dots represent forged images.

Comparison with the SoTA Methods

We compare our method with ManTraNet (Wu, AbdAlmageed, and Natarajan 2019), SPAN (Hu et al. 2020), MVSS-Net(Chen et al. 2021c), CAT-Net (Kwon et al. 2022), PSCC-Net (Liu et al. 2022) , HiFi-Net (Guo et al. 2023) and DiffForensics (Yu et al. 2024). The retrained results of comparison methods are reported from DiffForensics.

Comparisons on Localization. We show the pixel level localization performance in Table 1. The results show our method achieved superior performance across all datasets, particularly on the DGM forgery datasets. This could be attributed to MoENE's ability to handle complex and diverse tampering artifacts.

Comparisons on Detection. We show the detection results in Table 2. It can be seen that our method achieves the best accuracy performance and second best AUC and EER performance, demonstrating the effectiveness of FMAE. It is worth noting that IMD2020 is a real-world dataset with unknown forgery types. We achieved the best performance on IMD2020, which proves that our method can cope with tampering in unknown scenes.

Qualitative Comparisons. We provide predicted masks of different methods in Figure 5. It demonstrates that FMAE is not only able to locate the forged region more accurately but also able to form clearer boundaries, thanks to the natural image prior and multi-source forgery information.

Ablation Analysis

In this section, we conduct experiments to assess the impact of all design choices of our method. We initially address

Methods	Editing									DGM			Average		
	Casiav1.0			Columbia			IMD2020			AutoSplice					
	ACC ↑	AUC ↑	EER ↓	ACC ↑	AUC ↑	EER ↓	ACC ↑	AUC ↑	EER ↓	ACC ↑	AUC ↑	EER ↓	ACC ↑	AUC ↑	EER ↓
ManTraNet	.535	.546	.446	.496	.869	.219	.830	.698	.372	.614	.378	.586	.619	.623	.406
MVSSNet	.791	.937	.136	.664	.984	**.055**	.799	.661	.391	**.809**	.886	.191	.766	.867	.193
CAT-Net	.671	.690	.362	.755	.953	.115	.785	.684	.370	.699	.790	.296	.728	.779	.286
PSCC-Net	**.992**	**.999**	**.006**	.606	.981	.082	.821	.624	.425	.733	.877	.192	.788	.870	.176
HiFi-Net	.632	.717	.320	.532	.741	.917	**.826**	.523	.483	.618	.527	.457	.652	.627	.394
DiffForensics	.741	.991	.043	.895	.982	**.055**	.749	.740	.333	.696	**.951**	**.092**	.770	**.916**	**.131**
Ours	.834	.913	.153	**.912**	**.989**	.079	**.826**	**.767**	**.275**	.651	.940	.090	**.806**	.902	.149

Table 2: Image-level ACC, AUC and EER performance of image forgery detection. Our method achieves the highest ACC performance and the suboptimal AUC and EER performance.

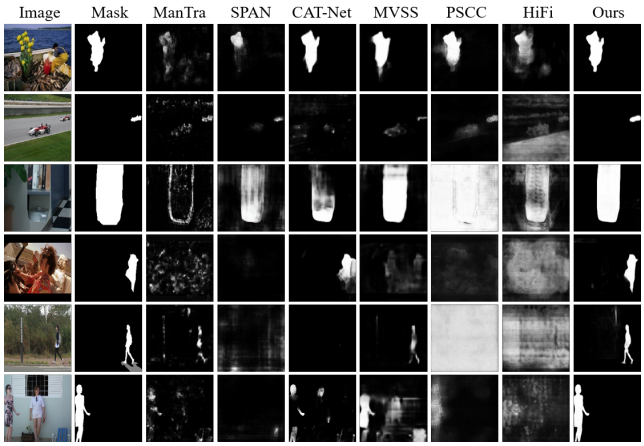


Figure 5: Visualization of the predicted manipulation mask by different methods.

three core designs of our method: Lottery Ticket Sparse Fine-tuning (LTSFT), 'Mixture of Experts' Noise Extractor (MoENE), and the input of forged information to the layer where the forged sensitive parameters are located (F2S). To evaluate the effectiveness of LTSFT, MoENE and F2S, we replace each of them separately from our method and evaluate the forgery localization performance on IMD2020 and AutoSplice, as shown in Table 3. We replace lottery ticket sparse fine-tuning by randomly unfreezing 70 % of parameters; remove MoENE from our network, and replace the present scheme F2S by feeding the forged information to the first four layers of the encoder.

The reduced performance of w/o LTSFT in comparison to the baseline is attributed to the fewer trainable parameters and their random selection. It can be seen that our three core designs all contribute to performance. In particular, when the Lottery Ticket Sparse Fine-tuning is removed, a certain degree of performance degradation can be seen. We unfreeze the encoder parameters to different degrees for comparison. It is worth noting that the parameters for unfreezing are selected based on the lottery ticket hypothesis. For a fair comparison, we both use the default adding noise strategy. It can be seen that when 70 % of the parameters are unfrozen, our model achieves the best performance on these two datasets. As the number of unfreezing parameters increases, the network fitting ability will be enhanced, but the generalization brought by the natural image prior will be attenuated. 70 % may be at the balance point between network generalization ability and fitting ability.

Set Up	Unfreeze	Components			IMD2020		AutoSplice	
		L	M	F	F1	AUC	F1	AUC
Baseline	100%	-	-	-	.384	.831	.435	.817
Different degrees of LTSFT	100%	+	+	+	.446	.866	.589	.849
	50%	+	+	+	.361	.806	.412	.834
	30%	+	+	+	.301	.832	.397	.805
w/o LTSFT	Randomly	-	+	+	.375	.865	.427	.807
w/o MoENE	70%	+	-	+	.429	.831	.560	.868
w/o F2S	70%	+	+	-	.445	.850	.607	.874
Full setup	70%	+	+	+	**.491**	**.894**	**.639**	**.950**

Table 3: Ablation studies on IMD2020 and AutoSplice. (L: LTSFT, M: MoENE, F: F2S.)

Conclusion

Our novel method, the Forgery Masked Autoencoder (FMAE), represents a groundbreaking shift in combating the rising threat of sophisticated image forgery. By prioritizing the universal traits of authentic images over specific tampering artifacts, FMAE excels in out-of-distribution forgery detection. The modification to the Masked Autoencoder framework preserves natural image features while seamlessly integrating multi-source forgery information. Leveraging the lottery ticket hypothesis during pre-training enables targeted fine-tuning of forgery-sensitive parameters, optimizing the model for detection and localization tasks. Additionally, the introduction of a 'mixture of experts' noise extractor enhances the model's resilience by aggregating forgery data from diverse sources. FMAE exhibits robustness against unseen forgeries, validating its efficacy in practical scenarios. Extensive experiments across diverse datasets confirm FMAE's superior accuracy and generalization capabilities compared to existing methods. As image forgery techniques evolve, FMAE stands as a resilient solution, highlighting the significance of universal image characteristics for IFDL.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62436008, 62422609 and 62276243.

References

Ansell, A.; Ponti, E.; Korhonen, A.; and Vulić, I. 2022. Composable Sparse Fine-Tuning for Cross-Lingual Transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1778–1796.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Bayar, B.; and Stamm, M. C. 2018. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 13(11): 2691–2706.

Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; and Sutskever, I. 2020. Generative pretraining from pixels. In *International conference on machine learning*, 1691–1703. PMLR.

Chen, T.; Frankle, J.; Chang, S.; Liu, S.; Zhang, Y.; Carbin, M.; and Wang, Z. 2021a. The lottery tickets hypothesis for supervised and self-supervised pre-training in computer vision models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16306–16316.

Chen, T.; Sui, Y.; Chen, X.; Zhang, A.; and Wang, Z. 2021b. A unified lottery ticket hypothesis for graph neural networks. In *International conference on machine learning*, 1695–1706. PMLR.

Chen, X.; Ding, M.; Wang, X.; Xin, Y.; Mo, S.; Wang, Y.; Han, S.; Luo, P.; Zeng, G.; and Wang, J. 2023. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 1–16.

Chen, X.; Dong, C.; Ji, J.; Cao, J.; and Li, X. 2021c. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14185–14193.

Chen, X.; Zhang, Z.; Sui, Y.; and Chen, T. 2021d. Gans can play lottery tickets too. *arXiv preprint arXiv:2106.00134*.

De Carvalho, T. J.; Riess, C.; Angelopoulou, E.; Pedrini, H.; and de Rezende Rocha, A. 2013. Exposing digital image forgeries by illumination color classification. *IEEE Transactions on Information Forensics and Security*, 8(7): 1182–1194.

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3539–3553.

Dong, J.; Wang, W.; and Tan, T. 2013. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*, 422–426. IEEE.

Frankle, J.; and Carbin, M. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*.

Fridrich, J.; and Kodovsky, J. 2012. Rich models for steganalysis of digital images. *IEEE Transactions on information Forensics and Security*, 7(3): 868–882.

Ge, C.; Fu, X.; He, P.; Wang, K.; Cao, C.; and Zha, Z.-J. 2024. Neuromorphic Event Signal-Driven Network for Video De-raining. In *AAAI*, volume 38, 1878–1886.

Ge, C.; Fu, X.; and Zha, Z.-J. 2022. Learning Dual Convolutional Dictionaries for Image De-raining. In *ACM MM*, 6636–6644.

Guan, H.; Kozak, M.; Robertson, E.; Lee, Y.; Yates, A. N.; Delgado, A.; Zhou, D.; Kheyrkhah, T.; Smith, J.; and Fiscus, J. 2019. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops*, 63–72. IEEE.

Guillaro, F.; Cozzolino, D.; Sud, A.; Dufour, N.; and Verdoliva, L. 2023. 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*, 20606–20615.

Guo, X.; Liu, X.; Ren, Z.; Grosz, S.; Masi, I.; and Liu, X. 2023. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3155–3165.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hsu, Y.-F.; and Chang, S.-F. 2006. Detecting image splicing using geometry invariants and camera characteristics consistency. In *2006 IEEE International Conference on Multimedia and Expo*, 549–552. IEEE.

Hu, J.; Liao, X.; Gao, D.; Tsutsui, S.; Qin, Z.; and Shou, M. Z. 2023. DeepfakeMAE: Facial Part Consistency Aware Masked Autoencoder for Deepfake Video Detection. *arXiv preprint arXiv:2303.01740*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; and Nevatia, R. 2020. SPAN: Spatial pyramid attention network for image manipulation localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 312–328. Springer.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jia, S.; Huang, M.; Zhou, Z.; Ju, Y.; Cai, J.; and Lyu, S. 2023. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 893–903.

Kniaz, V. V.; Knyaz, V.; and Remondino, F. 2019. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. *Advances in neural information processing systems*, 32.

Korus, P.; and Huang, J. 2016. Multi-scale analysis strategies in PRNU-based tampering localization. *IEEE Transactions on Information Forensics and Security*, 12(4): 809–824.

Kwon, M.-J.; Nam, S.-H.; Yu, I.-J.; Lee, H.-K.; and Kim, C. 2022. Learning JPEG compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8): 1875–1895.

Le, T.-N.; Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2021. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10117–10127.

Li, D.; Zhu, J.; Fu, X.; Guo, X.; Liu, Y.; Yang, G.; Liu, J.; and Zha, Z.-J. 2025. Noise-Assisted Prompt Learning for Image Forgery Detection and Localization. In *European Conference on Computer Vision*, 18–36. Springer.

Li, D.; Zhu, J.; Wang, M.; Liu, J.; Fu, X.; and Zha, Z.-J. 2023. Edge-Aware Regional Message Passing Controller for Image Forgery Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8222–8232.

Liang, Z.; Wei, F.; Jie, Y.; Qian, Y.; Hao, Z.; and Han, B. 2023. Prompts can play lottery tickets well: Achieving lifelong information extraction via lottery prompt tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 277–292.

Lin, X.; Wang, S.; Deng, J.; Fu, Y.; Bai, X.; Chen, X.; Qu, X.; and Tang, W. 2023. Image manipulation detection by multiple tampering traces and edge artifact enhancement. *Pattern Recognition*, 133: 109026.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive Spatio-Channel Correlation Network for Image Manipulation Detection and Localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7505–7517.

Ma, X.; Du, B.; Liu, X.; Hammadi, A. Y. A.; and Zhou, J. 2023. Iml-vit: Image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*.

Nath, P. S.; Gandhi, H. K.; and Chouhan, R. 2021. Quantifying image naturalness using differential curvelet features. In *2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 1–6. IEEE.

Novozamsky, A.; Mahdian, B.; and Saic, S. 2020. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 71–80.

Panda, A.; Isik, B.; Qi, X.; Koyejo, S.; Weissman, T.; and Mittal, P. 2024. Lottery ticket adaptation: Mitigating destructive interference in llms. *arXiv preprint arXiv:2406.16797*.

Peng, L.; Cao, Y.; Sun, Y.; and Wang, Y. 2024. Lightweight Adaptive Feature De-drifting for Compressed Image Classification. *IEEE Transactions on Multimedia*.

Ren, W.; Ma, L.; Zhang, J.; Pan, J.; Cao, X.; Liu, W.; and Yang, M.-H. 2018. Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3253–3261.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A.; and Bottou, L. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).

Wu, Y.; AbdAlmageed, W.; and Natarajan, P. 2019. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9543–9552.

Yu, H.; Edunov, S.; Tian, Y.; and Morcos, A. S. 2019. Playing the lottery with rewards and multiple languages: lottery tickets in rl and nlp. *arXiv preprint arXiv:1906.02768*.

Yu, Z.; Ni, J.; Lin, Y.; Deng, H.; and Li, B. 2024. DiffForensics: Leveraging Diffusion Prior to Image Forgery Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12765–12774.

Zhang, F.; Liu, J.; Xie, J.; Zhang, Q.; Xu, Y.; and Zha, Z.-J. 2024. ESCNet: Entity-enhanced and Stance Checking Network for Multi-modal Fact-Checking. In *Proceedings of the ACM on Web Conference 2024*, 2429–2440.

Zhang, F.; Liu, J.; Zhang, Q.; Sun, E.; Xie, J.; and Zha, Z.-J. 2023. ECENet: Explainable and Context-Enhanced Network for Muti-modal Fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1231–1240.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6881–6890.

Zhou, P.; Chen, B.-C.; Han, X.; Najibi, M.; Shrivastava, A.; Lim, S.-N.; and Davis, L. 2020. Generate, segment, and refine: Towards generic manipulation segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13058–13065.

Zhou, P.; Han, X.; Morariu, V. I.; and Davis, L. S. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1053–1061.

Zhu, J.; Li, D.; Fu, X.; Yang, G.; Huang, J.; Liu, A.; and Zha, Z.-J. 2024. Learning Discriminative Noise Guidance for Image Forgery Detection and Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7739–7747.