M^{2}RL-Net: Multi-View and Multi-Level Relation Learning Network for Weakly-Supervised Image Forgery Detection

Jiafeng Li[1], **Ying Wen**[1*], **Lianghua He**[2]

[1]School of Communication and Electronic Engineering, East China Normal University
[2]Department of Computer Science and Technology, Tongji University
51205904113@stu.ecnu.edu.cn, ywen@cs.ecnu.edu.cn, helianghua@tongji.edu.cn

Abstract

As digital media manipulation becomes increasingly sophisticated, accurately detecting and localizing image forgeries with minimal supervision has become a critical challenge. Existing weakly supervised image forgery detection (W-IFD) methods often rely on convolutional neural networks (CNNs) and limited exploration of internal relationships, leading to poor detection and localization performance with only image-level labels. To address these limitations, we introduce a novel Multi-View and Multi-Level Relation Learning Network (M^{2}RL-Net) for W-IFD. M^{2}RL-Net effectively identifies forged images using only image-level annotations by exploring relationships between different views and hierarchical levels within images. Specifically, M^{2}RL-Net achieves patch-level self-consistency learning (PSL) and feature-level contrastive learning (FCL) across different views, facilitating more generalized self-supervised learning of forgery features. In detail, PSL employs self-supervised learning to distinguish consistent and inconsistent regions within images, enhancing its ability to accurately locate tampered areas. FCL utilizes feature-level self-view and multi-view contrastive learning to differentiate between genuine and tampered image features, thereby improving the recognition of authentic and manipulated content across different views. Extensive experiments on various datasets demonstrate that M^{2}RL-Net outperforms existing weakly supervised methods in detection and localization accuracy. This research sets a new benchmark for W-IFD and lays a robust foundation for future studies in this field.

Introduction

Digital images play a crucial role in disseminating information and documenting reality. However, the proliferation of advanced editing tools and models, such as Diffusion and GANs (Franceschi et al. 2024; Rombach et al. 2022) has made image manipulation increasingly accessible, challenging the notion of "seeing is believing." The edited or manipulated images can fabricate fake news, perpetrate academic fraud, and facilitate illegal activities, thereby threatening social stability. Detecting such manipulation is difficult, making it imperative to advance image forgery detection to verify image authenticity and spot the tampered areas. Most existing deep learning (DL)-based methods (Zhang et al. 2024a; Zhu et al. 2024; Yu et al. 2024; Ma et al. 2023) for image forgery detection mainly focus on fully supervised learning to extract tampered artifact features, requiring extensive pixel-level annotations. While effective to a certain extent, these methods face high annotation costs and scalability issues. As image processing technologies evolve, traditional methods often fail to adapt to new types of manipulations, especially those generated using advanced AI techniques (Qu et al. 2024; Zhang, Li, and Chang 2024; Zhang et al. 2024b). Moreover, DL-based image forgery detection methods typically perform well on training datasets but exhibit significant performance degradation on unknown images, limiting their effectiveness in real-world applications. While pixel-level annotations provide full supervision for differentiating between authentic and tampered regions, their high cost restricts the number of training images.

Given these challenges, we propose a weakly-supervised image forgery detection (W-IFD) method, called Multi-View Multi-Level Relation Learning Network (M^{2}RL-Net), which only requires binary image-level labels to localize the manipulated area, eliminating the need for detailed pixel-level masks during training. In W-IFD, we commonly focus on identifying inconsistencies between normal (authentic) and abnormal (tampered) areas within an image. We posit that by investigating how features in manipulated regions relate to those in surrounding areas across various views and levels of detail, we can overcome the limitations associated with relying exclusively on image-level annotations. Specifically, M^{2}RL-Net leverages multi-view feature representations (MFR) and employs two relation learning strategies: patch-level self-consistency learning (PSL) and feature-level contrastive learning (FCL) to improve its generalization in detecting and localizing complex tampering scenarios.

MFR captures feature representations from both RGB and noise views, where the RGB stream is designed to detect visually apparent tampering artifacts, while the noise stream focuses on identifying distribution inconsistencies between tampered and authentic regions to identify semantic-agnostic artifacts. PSL explores internal patch-level similarities to enforce consistency constraints within ambiguous areas of forgery localization. This self-supervised technique utilizes generated pseudo-localization regions to correct potential false positives, precisely estimate inconsistent manipulation regions, and mitigate overfitting. FCL asserts that

*Corresponding author: Ying Wen.

pixels of the same class have similar representations in the feature space, thus we can extract category knowledge prototypes for both authentic and forged classes and extend it for contrastive learning under self-view and multi-view conditions using supervision from reliable pixels, differentiating between authentic and tampered features.

Overall, M^{2}RL-Net integrates MFR, PSL, and FCL to facilitate relation learning at different levels, enhancing the detection of inconsistency manipulations and identifying manipulative pattern features across different views. Our method enhances the model's pixel-level detection of forged regions and improves its generalization capabilities.

In summary, our contributions are listed as follows:

- We propose a novel end-to-end weakly-supervised image forgery detection method called M^{2}RL-Net, which reduces dependency on extensive labeled datasets while achieving high accuracy in detecting and localizing manipulated regions.
- We introduce Patch-Level Self-Consistency Learning (PSL) to leverage inherent patch-level similarity for self-consistency learning, enhancing generic image manipulation detection.
- We propose Multi-view Feature Representation (MFR) and Feature-Level Contrastive Learning (FCL) to learn and extract prototypes of genuine and forged knowledge from various feature representations, utilizing reliable pixels for self-view and multi-view contrastive learning, cultivating more discriminative genuine and forged features, thereby enhancing general detection capabilities.
- We have extensively validated our methods across multiple benchmark datasets, demonstrating superior performance in both detection accuracy and manipulation localization compared to existing fully-supervised and weakly-supervised methods.

Related Work

Image Forgery Detection Early methods design handcrafted features to identify specific manipulations. Given the diverse and often unknown nature of real-world editing, recent studies focus on practical general tampering detection (Triaridis and Mezaris 2024; Guillaro et al. 2023). These methods detect forgeries by identifying multiple traces like JPEG artifacts, edge inconsistencies, noise patterns, and camera model fingerprints. Researchers (Li et al. 2024; Zeng et al. 2024; Guo et al. 2023; Zhou et al. 2018) enhance detection by exploring clues beyond RGB views, converting images into noise views with fixed or learnable filters like SRM and learnable convolutions to highlight artifacts. Some studies (Kwon et al. 2022, 2021; Wang et al. 2022) incorporate DCT coefficients alongside RGB to detect compression artifacts. Recently, contrastive learning has been applied to image forgery detection, with methods like (Wu, Chen, and Zhou 2023; Lou et al. 2024; Niloy, Bhaumik, and Woo 2023) using contrastive loss to differentiate authentic and forged pixel embeddings. These studies rely on annotated ground truth labels for pixel-level or block-level contrastive learning in fully supervised settings, which has not been explored in weakly supervised contexts. Fully supervised methods, though effective, are time-consuming, result in large training datasets, and increase false positives. Besides, new editing techniques like language-driven tampering do not always produce pixel-level masks. Weakly supervised image forgery detection has been successively proposed to address the limitations of fully supervised methods.

Weakly-supervised Learning Weakly-supervised learning strategies are designed to alleviate the extensive labeling effort required by fully-supervised systems. This paradigm enables the prediction of fine-grained labels using coarse or incomplete supervision, such as using image-level labels to predict segmentation masks. Given the relatively limited research on weakly-supervised image forgery detection (W-IFD), this paper references the latest techniques WSCL (Zhai et al. 2023) and EdgeCAM (Zhou et al. 2024) and compares them with weakly-supervised segmentation algorithms. Our method demonstrates significant advantages at both the image and pixel levels, enabling effective generalization across varied and unseen manipulation scenarios.

Methodology

In this section, we present the proposed Multi-View and Multi-Level Relation Learning Network (M^{2}RL-Net) framework designed to tackle the challenges in weakly-supervised image forgery detection (W-IFD).

Problem Definition

In this paper, we try to explore the internal properties of the forgery detection network to localize partially manipulated images when only image-level labels are available. For the image forgery detection network $Net(\cdot)$, given an input image $I \in \mathbb{R}^{H\times W\times 3}$, we have an image-level manipulation label $y_I \in \{0, 1\}$, where 0 indicates authentic images and 1 indicates fake images. The network generates a prediction map $m = Net(I) \in \mathbb{R}^{H\times W}$, from which we derive the image-level prediction score $y = Pooling(m)$. During training, we only utilize image-level labels for supervision. At inference, we not only identify whether an image is authentic or forged but generate a predicted map m to localize forged regions at the pixel level within manipulated images.

Overview of M^{2}RL-Net

As illustrated in Fig. 1, we present an overview of M^{2}RL-Net, which comprises three main components: Multi-view Feature Representation (MFR), Patch-level Self-Consistency Learning (PSL), and Feature-level Contrastive Learning (FCL). The framework aims to fully exploit the forgery artifacts representation from diverse views to provide robust cues for manipulation detection and leverage multi-level relation learning to discern manipulative patterns across different levels for weakly-supervised image forgery detection and localization.

Multi-View Feature Representation (MFR)

Tampering operations often disrupt the natural noise distribution between the source and target images. Thus we posit that mining noise inconsistencies can provide robust

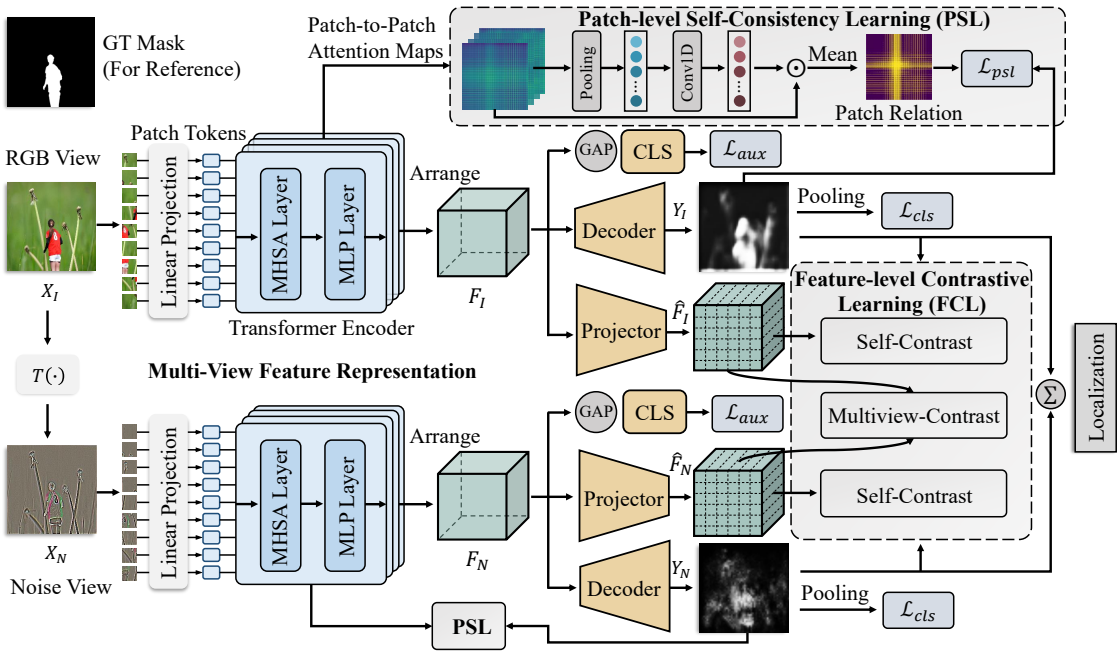

Figure 1: Overall architecture of our M^{2}RL-Net framework, which mainly consists of: a multi-view feature representation, a decoder and feature projector for two relation learning (PSL and FCL), and a classification layer (CLS) for auxiliary tasks.

artifact cues for forgery detection. Specifically, we employ a dual-stream structure to fully exploit and integrate clues from both RGB and noise perspectives during the multi-view feature representation (MFR) stage. Initially, the input RGB image X_I is transformed into a noise view representation $X_N = T(X_I)$. The transformation $T(\cdot)$ utilizes the learnable constrained convolutional layer (Bayar and Stamm 2018), which can reveal low-level tampering inconsistencies. Subsequently, the input image X_I and noise map X_N are processed through the feature encoding pipeline. We explore Transformer-based models (Xie et al. 2021) that achieved astonishing performance on visual classification benchmarks. Specifically, X_I and X_N are divided into N^2 patches of size $\mathcal{P}$, and these patches are flattened into embeddings and linearly mapped into N^2 patch tokens. These tokens as the transformer encoder's input $z_{in} \in \mathbb{R}^{N^2 \times D}$ are fed into the feature encoding stage, where D is the dimension of input tokens. The transformer encoder consists of L encoding layers internally. Each layer, ℓ, comprises two sub-layers: a multi-head self-attention (MHSA) layer (Vaswani et al. 2017) and a multilayer perceptron (MLP) layer. Within each encoding layer, we input tokens z_{in} and receive z_{out}. The z_{out} from one layer becomes the z_{in} for the subsequent encoder layer, iterating for L iterations.

$$\mathbf{y}^{\ell} = \mathrm{MHSA}\left(\mathrm{LN}\left(\mathbf{z}^{\ell-1}\right)\right) + \mathbf{z}^{\ell-1} \quad (1)$$

$$\mathbf{z}^{\ell} = \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{y}^{\ell}\right)\right) + \mathbf{y}^{\ell} \quad (2)$$

where LN is layer normalization (Ba, Kiros, and Hinton 2016) and MLP consists of two linear projections separated by GeLU (Hendrycks and Gimpel 2016) non-linearity.

After going through the L transformer layers, these encoded tokens $\mathbf{z}_{last}$ are then arranged as the final extracted features $F_n (n \in \{I, N\})$. After the MFR stage, we can acquire discriminative feature representations of input images.

Patch-Level Self-Consistency Learning (PSL)

Prior works (Ru et al. 2022; Zhu et al. 2023) have indicated in each transformer block of the standard Transformer Encoder, there exists a Multi-Head Self Attention (MHSA) layer that calculates the similarity between queries and keys from different patch tokens, referred to as Patch Attention Maps. Different from the weakly-supervised WSCL method (Zhai et al. 2023), which computes the pair-wise similarity of extracted features, our transformer-based method directly uses the Patch Attention Maps from intermediate layers for self-consistency learning. Consequently, as shown in Fig. 1, we propose Patch-Level Self-Consistency Learning (PSL) to leverage the Patch Attention Maps in the transformer encoder to identify inconsistencies between image patches and differentiate feature representations of authentic and tampered patches. Concretely, within the transformer encoder in MFR, we obtain the Patch Attention Maps of N^2 different patch tokens relative to themselves from the L MHSA layers, which is denoted as $A \in \mathbb{R}^{(L \times H) \times N^2 \times N^2}$, where L is the MHSA layers, H is the number of attention heads, N^2 is the number of patch tokens. $A^{(i,j),(k,l)}$ represents the consistency between the patch token at position (i, j) and another patch token at position (k, l), with higher values indicating greater consistency.

To combine the representation of all attention heads in MHSA layers, it typically directly averages the self-attention maps of different heads in the same layer and then sums them by different layers. This mean-sum approach to the Patch Attention Maps tends to introduce more interference

to the activation of forgery regions. Thus, we propose utilizing an adaptive learning module to recalibrate the importance of different attention heads. As shown in Fig. 1, first, we get the Patch Attention Maps $A \in \mathbb{R}^{(L\times H)\times N^2\times N^2}$ and aggregate the corresponding global context information $W \in \mathbb{R}^{(L\times H)\times 1}$ by applying pooling across the heads. Subsequently, we utilize a fast 1D convolution (Conv1D) of size k to efficiently facilitate interaction among the different attention heads as follows:

$$W^{'} = \text{Conv1D}(\text{Pooling}(A)) \tag{3}$$

Finally, we multiply the interacted weights $W^{'} \in \mathbb{R}^{(L\times H)\times 1}$ back to the Patch Attention Maps A and apply a mean operation to obtain a comprehensive Patch Relation Map $\hat{A} \in \mathbb{R}^{N^2\times N^2}$ as follows:

$$\hat{A} = \frac{1}{LH}\sum_{i=1}^{L\cdot H} W_i^{'} \cdot A_i \tag{4}$$

During training, the Patch Relation Map $\hat{A}$ is typically coarse and inaccurate due to the absence of explicit constraints. With the approximate pseudo-forgery location map $Y_n (n \in \{I, N\})$, the pseudo-localization label M_n is derived. Specifically, for M_n, if the pixels (i, j) and (k, l) share the same consistency, their patch-level consistency similarity is set to 1; otherwise, it is set to 0, as follows:

$$C_{(i,j),(k,l)} = \begin{cases} 1, & \text{if } M_{ij} = M_{kl} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

The patch self-consistency loss term $\mathcal{L}_{psl}$ is formalized as:

$$\mathcal{L}_{psl} = \frac{1}{N^2}\sum_{i,j,k,l} |\text{sigmoid}(\hat{A}_{(i,j),(k,l)}) - C_{(i,j),(k,l)}|, \tag{6}$$

Overall, $\mathcal{L}_{psl}$ imposes consistency constraints in confusing regions, forming a self-supervised paradigm. PSL leverages self-consistency learning to detect inconsistent image patches and differentiate features between real and tampered patches, enhancing the final more reliable prediction.

Feature-Level Contrastive Learning (FCL)

Relying solely on PSL remains challenging for distinguishing between real and forged regions under weak supervision, which depends only on image-level labels. To address this, we introduce a Feature-Level Contrastive Learning (FCL) module that explores feature-level relationships, ensuring that pixels of the same class have similar representations in the feature space, thereby enhancing detection performance.

FCL operates by extracting category-specific knowledge prototypes for both real and forged classes, which serve as representative embeddings, helping to reduce noise from erroneous pseudo-labels in weakly supervised settings. By introducing corresponding positive prototype pixels, FCL drives negative prototype pixels apart within the projected feature space, enabling contrastive learning to generate more discriminative feature embeddings for each pixel.

Initially, in the absence of labeled pixels for image tagging, we depend on trustworthy pseudo-label maps to guide

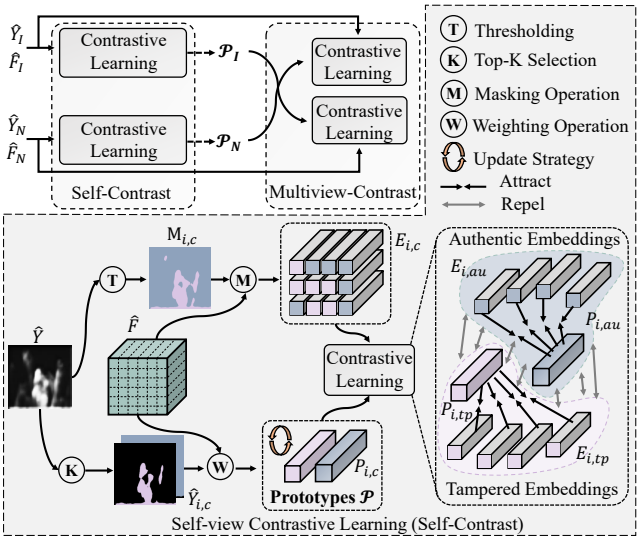

Figure 2: Illustration of the proposed FCL. Self-contrast and multi-view contrast are formed.

FCL. From these maps, we select reliable pixels for each class to compute their class prototypes. The process involves pixel-level contrastive learning between the class features and prototypes within the feature space, enabling us to obtain discriminative feature embeddings for each pixel. Following (Du et al. 2022), these pseudo-label maps $\hat{Y}_n \in \mathbb{R}^{1\times H\times W}$ are derived by selecting class prediction maps Y_n, applying a ReLU function and normalizing the values to the $[0, 1]$ range as estimates of the probability of belonging to the forged class. Subsequently, we apply a feature projector to the feature representations F_n to produce pixel-level projected feature embeddings $\hat{F}_n \in \mathbb{R}^{128\times H\times W}$, which includes 1×1 convolutional layers and ReLU activation.

As illustrated in Fig. 2, with the pseudo-label maps $\hat{Y}_n$ and projected feature embeddings $\hat{F}_n$, we estimate the prototypes, defined as the discriminative feature embeddings for each class, by calculating them from the most reliable pixels in the pseudo-label maps. Specifically, for pixels in $\hat{Y}_n$ assigned to different classes, we select the top-K most confident pixels to estimate prototypes. We then use the reliable class pseudo-labels to categorize the pixel-level projected feature embeddings $\hat{F}_n$. A weighted averaging is performed on all pseudo-labeled features $\hat{F}_n$ to derive the prototypes $\mathcal{P}_c$ as follows:

$$\mathcal{P}_c = \frac{\sum_{i\in\Omega_c} \hat{Y}_{i,c}\hat{F}_i}{\sum_{i\in\Omega_c} \hat{Y}_{i,c}} \tag{7}$$

Here, $\hat{Y}_c$ represents the pseudo-label map for class c, and Ω_c denotes the set of top-K pixels in the pseudo-label map for class c. The hyperparameter K determines the range of pixels used to compute the prototype, where a smaller K indicates higher confidence in the prototype estimation.

To obtain class-specific pixel feature embeddings E_c, we use the pseudo-label maps $\hat{Y}$ as a probability map for pixel-level classes. By applying thresholding ρ, where pixels with

values above or equal to ρ are marked as forged, and those below ρ are marked as genuine, we generate class masks $M_{i,c}$. Consequently, the projected feature embeddings $\hat{F}_i$ are divided into tampered embeddings $E_{i,tp}$ and authentic embeddings $E_{i,au}$ as shown in the right part of Fig. 2.

Finally, the similarity between class feature embeddings and their positive prototypes ($E_{i,tp}$ and $P_{i,tp}$, $E_{i,au}$ and $P_{i,au}$) should be maximized, while the similarity between class features and negative prototypes ($E_{i,tp}$ and $P_{i,au}$, $E_{i,au}$ and $P_{i,tp}$) should be minimized, as illustrated in Fig. 2. In this study, the contrastive learning (CL) between class feature embeddings and prototypes is formulated as follows:

$$\mathrm{CL}(\boldsymbol{E}_i; \mathcal{P}; \boldsymbol{M}_i) = -\log \frac{\exp(\boldsymbol{E}_i \cdot \mathcal{P}_{\boldsymbol{M}_i}/\tau)}{\sum_{c \in C} \exp(\boldsymbol{E}_i \cdot \mathcal{P}_c/\tau)} \tag{8}$$

where $\boldsymbol{M}_i$ is the pseudo-label of pixel i, which determines the positive prototype $\mathcal{P}_{\boldsymbol{M}_i}$. The temperature coefficient τ is set to 0.1, and C represents the set of both forged and authentic classes.

Self-Contrast. For self-view contrastive learning, the comparison is conducted within a single view of each image. For a pixel i with pseudo-label Y_i, self-view contrastive learning extracts the prototype $\mathcal{P}$ from the current view and conducts contrastive learning between pixel class features and prototypes as follows:

$$\mathcal{L}_{\mathrm{self}} = \frac{1}{|V|} \sum_{i \in V} \left(\mathrm{CL}(\mathbf{E}_{\mathrm{I}}, \mathcal{P}_{\mathrm{I}}, \mathbf{M}_{\mathrm{I}}) + \mathrm{CL}(\mathbf{E}_{\mathrm{N}}, \mathcal{P}_{\mathrm{N}}, \mathbf{M}_{\mathrm{N}})\right) \tag{9}$$

where V indicates the entire image and $|\cdot|$ represents the cardinality. I and N denote the input RGB and noise view.

However, due to the lack of precise pixel annotations in weak supervision settings, the pseudo-label $\hat{Y}_i$ assigned to pixel i may be inaccurate, leading to inaccurate prototypes $\mathcal{P}$ and thus incorrect contrast between category features and prototypes. We address this issue from two aspects. On one hand, to fully utilize reliable forged and real pixels across the entire dataset without being constrained by the batch size, we propose a prototype update strategy similar to Exponential Moving Average (EMA) to update the prototypes $\mathcal{P}$, thereby learning more generalized prototypes for forged and real classes. The process is as follows:

$$\mathcal{P}_k \leftarrow \partial \mathcal{P}_{k-1} + (1 - \partial)\mathcal{P}_k \tag{10}$$

where $\mathcal{P}_{k-1}$ and $\mathcal{P}_k$ are the prototypes at the $(k-1)$-th and k-th iterations during training, and ∂ is the update weight.

Multiview-Contrast. On the other hand, considering the consistency of features from the two streams towards manipulation characteristics, we construct multi-view contrastive learning, using prototypes from other views to provide supervision signals for the current view and vice versa, as shown in Fig. 2. Given a pixel i with pseudo-label $\hat{Y}_v$ and projected feature embeddings $\hat{F}_v$, we regularize the current view using prototypes $\mathcal{P}_v (v \in \{I, N\})$ from another view. The multi-view contrastive loss is calculated as:

$$\mathcal{L}_{\mathrm{multi}} = \frac{1}{|V|} \sum_{i \in V} \left(\mathrm{CL}(\mathbf{E}_{\mathrm{I}}, \mathcal{P}_{\mathrm{N}}, \mathbf{M}_{\mathrm{I}}) + \mathrm{CL}(\mathbf{E}_{\mathrm{N}}, \mathcal{P}_{\mathrm{I}}, \mathbf{M}_{\mathrm{N}})\right) \tag{11}$$

Dataset	#Au	#Tp	#Spli.	#Cpmv.	#Inpa.	#Ps.
CASIAv2	7,491	5,063	3,235	1,828	-	-
CASIAv1	800	920	459	461	-	-
Columbia	183	180	180	-	-	-
COVER	100	100	-	100	-	-
NIST16	-	564	288	68	208	-
IMD20	2010	2010	-	-	-	2010

Table 1: Public dataset details. Au, Tp, Cpmv., Spli., Inpa. and Ps. are abbreviations for Authentic, Tampered, Copy-move, Splicing, Inpainting, and Photoshop, respectively.

The FCL module integrates self-view and multi-view contrastive learning, which can be expressed as:

$$\mathcal{L}_{\mathrm{fcl}} = \mathcal{L}_{\mathrm{self}} + \mathcal{L}_{\mathrm{multi}} \tag{12}$$

Optimization and Inference

As illustrated in Fig. 1, our framework consists of four loss terms: an auxiliary classification loss $\mathcal{L}_{\mathrm{aux}}$, a binary classification loss $\mathcal{L}_{\mathrm{cls}}$, a patch-level self-consistency learning (PSL) loss $\mathcal{L}_{\mathrm{psl}}$, and a feature-level contrastive learning (FCL) loss $\mathcal{L}_{\mathrm{fcl}}$. For the classification loss, following common practice, we employ a global average pooling (GAP) layer followed by a classification layer (CLS) to compute the class probability vector. $\mathcal{L}_{\mathrm{aux}}$ is applied using the Binary Cross-Entropy (BCE) criterion. Additionally, we utilize global max pooling on the predicted pseudo mask to generate image-level predictions, where the BCE loss serves as the binary classification loss $\mathcal{L}_{\mathrm{cls}}$. The PSL loss $\mathcal{L}_{\mathrm{psl}}$ and FCL loss $\mathcal{L}_{\mathrm{fcl}}$ are specifically defined in Eq. 6 and Eq. 12, respectively. The overall loss is the weighted sum of these loss terms:

$$\mathcal{L} = \mathcal{L}_{\mathrm{aux}} + \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{psl}} + \mathcal{L}_{\mathrm{fcl}} \tag{13}$$

The final image-level prediction is obtained by weighted averaging the predictions from RGB and noise streams, and the prediction map is the ensemble localization map.

Experiments

Experimental Settings

Datasets. To ensure a fair basis for comparison, we adhered to established test protocols outlined in prior studies (Dong et al. 2022; Zhai et al. 2023). The training dataset utilized in our experiments is CASIAv2 (Dong, Wang, and Tan 2013). For in-dataset evaluation, we employ the CASIAv1 dataset, while for cross-dataset evaluation, we utilize the Columbia (Hsu and Chang 2006), COVER (Wen et al. 2016), NIST16 (Guan et al. 2019), and IMD20 (Novozamsky, Mahdian, and Saic 2020) datasets as testing datasets. Our approach relies solely on image-level labels without any fine-tuning on the target datasets. Detailed information on the datasets is provided in Tab. 1.

Evaluation Metrics. In our experiments, we evaluate image-level manipulation detection using specificity (Spe.), sensitivity (Sen.), their F1 score (I-F1), and the area under the ROC curve (AUC). For pixel-level localization, we

Method	CASIAv1				Columbia				COVER				IMD20				AVG	
	AUC	Spe.	Sen.	I-F1	AUC	Spe.	Sen.	I-F1	AUC	Spe.	Sen.	I-F1	AUC	Spe.	Sen.	I-F1	AUC	I-F1
Unsupervised methods.																		
NOI1	0.500	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000
CFA1	0.482	0.000	1.000	0.000	0.344	0.000	1.000	0.000	0.525	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000
Fully-supervised methods.																		
H-LSTM	0.498	0.0	0.997	0.000	0.506	0.001	1.000	0.002	0.500	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.515	0.001
ManTra-Net	0.141	0.000	1.000	0.000	0.701	0.000	1.000	0.000	0.491	0.000	1.000	0.000	0.719	0.000	1.000	0.000	0.513	0.000
RRU-Net	0.507	0.006	0.994	0.001	0.497	0.000	1.000	0.000	0.495	0.000	1.000	0.000	0.512	0.000	1.000	0.000	0.503	0.000
CR-CNN	0.766	0.224	0.930	0.361	0.783	0.246	0.961	0.392	0.566	0.070	0.967	0.131	0.617	0.112	0.936	0.200	0.683	0.271
GSR-Net	0.502	0.011	0.994	0.022	0.502	0.011	1.000	0.022	0.515	0.000	1.000	0.000	0.505	0.008	0.998	0.014	0.506	0.019
SPAN	0.500	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000
CAT-Net	0.630	0.328	0.762	0.459	0.849	0.373	0.782	0.505	0.572	0.093	0.902	0.169	0.721	0.132	0.872	0.229	0.693	0.157
FCN+DA	0.796	0.844	0.717	0.775	0.762	0.322	0.950	0.481	0.541	0.100	0.900	0.180	0.746	0.030	0.981	0.182	0.711	0.404
MVSS-Net	0.937	0.988	0.615	0.758	0.980	1.000	0.669	0.802	0.731	0.940	0.140	0.244	0.656	0.915	0.220	0.355	0.826	0.534
Weakly-supervised methods.																		
MIL-FCN	0.647	0.538	0.569	0.553	0.807	0.220	0.732	0.338	0.542	0.062	0.793	0.115	0.578	0.116	0.886	0.205	0.644	0.303
MIL-FCN+WSCL	0.829	0.795	0.690	0.738	0.920	0.519	0.983	0.680	0.584	0.440	0.714	0.544	0.733	0.221	**0.966**	0.360	0.766	0.580
Araslanov	0.642	0.458	0.542	0.496	0.773	0.127	0.746	0.140	0.560	0.077	0.746	0.140	0.665	0.126	0.832	0.219	0.600	0.270
Araslanov+WSCL	0.796	0.638	0.726	0.679	0.917	0.324	0.948	0.483	0.591	0.220	**0.838**	0.348	0.701	0.193	0.872	0.316	0.751	0.456
EdgeCAM	0.836	**0.928**	0.713	0.806	0.897	0.776	0.694	0.733	**0.729**	0.470	0.270	0.343	0.642	0.582	0.648	**0.613**	0.776	0.624
WSCL(re-trained)	0.778	0.875	0.622	0.727	0.921	0.909	0.961	0.909	0.570	0.450	0.670	0.538	0.612	0.879	0.302	0.449	0.720	0.655
Ours	**0.948**	0.908	**0.827**	**0.866**	**0.999**	**0.951**	**1.000**	**0.975**	0.716	**0.860**	0.473	**0.610**	**0.827**	**0.990**	0.415	0.585	**0.862**	**0.762**

Table 2: Comparison of state-of-the-art methods for image-level manipulation detection across multiple datasets, evaluated by AUC, specificity (Spe.), sensitivity (Sen.), and image-level F1 score (I-F1). The best and second-best results are highlighted in **boldface** and underlined, respectively.

compute precision, recall and their F1 score (P-F1) on tampered images. We also use the combined F1 score (C-F1) for overall performance assessment, applying a uniform decision threshold of 0.5 for F1 computations.

Implementation Details. The proposed M^{2}RL-Net is implemented with PyTorch and trained on four NVIDIA RTX 3090 GPUs. Following (Zhai et al. 2023), all input images are resized to 224×224 and augmented by commonly used cropping and flipping. Segformer (Xie et al. 2021) is adopted as the transformer encoder for both views, which is pretrained on the ImageNet dataset, while the classifier, decoder, and projector are initialized using random weights. We use the AdamW optimizer (Loshchilov and Hutter 2019) with a batch size of 32 and an initial learning rate of $1e^{-4}$. The entire model is trained for 60 epochs. In FCL, the threshold ρ is set to 0.5 and the update weight ∂ to 0.1.

Comparison with State-of-the-Art

In this section, we compare M^{2}RL-Net's image-level detection and pixel-level localization performance with 15 existing methods (unsupervised: CFA1 (Ferrara et al. 2012), NOI1 (Mahdian and Saic 2009); fully supervised: H-LSTM (Bappy et al. 2019), ManTra-Net (Wu, AbdAlmageed, and Natarajan 2019), RRU-Net (Bi et al. 2019), CR-CNN (Yang et al. 2020), GSR-Net (Zhou et al. 2020), SPAN (Hu et al. 2020), CAT-Net (Kwon et al. 2022), FCN+DA (Chen et al. 2021), MVSS-Net (Dong et al. 2022); weakly supervised: MIL-FCN (Pathak et al. 2015), Araslanov (Araslanov and Roth 2020), WSCL (Zhai et al. 2023)), EdgeCAM (Zhou et al. 2024) in both in-dataset and cross-dataset setups.

Image-level Detection Results. Tab. 2 presents the results of the image manipulation detection. Compared to unsupervised and fully supervised methods, M^{2}RL-Net has competitive AUC and Image-level F1 (I-F1) scores, meaning our method outperforms the counterparts on both in-and-cross datasets. For instance, our method achieve a 42.7% improvement in average I-F1 compared to the supervised algorithm MVSS-Net. According to both sensitivity (Sen.) and specificity (Spe.) metrics, our method can learn from the authentic and obtain higher specificity, and thus lower false alarm rate, on all test sets. Furthermore, our method outperforms the equivalent weakly-supervised algorithm, WSCL, with a 19.7% improvement in AUC and a 16.3% boost in I-F1. Such results highlight the effectiveness of our method, especially the generalization ability and false detection rate in authentic images.

Method	P-F1					
	CASIAv1	Columbia	COVER	IMD20	NIST16	AVG
Unsupervised methods.						
NOI1	0.157	0.311	0.205	0.124	0.089	0.190
CFA1	0.140	0.320	0.188	0.111	0.106	0.188
Fully-supervised methods.						
H-LSTM	0.154	0.130	0.163	0.195	0.354	0.176
ManTra-Net	0.155	0.364	0.286	0.122	0.000	0.185
RRU-Net	0.225	0.452	0.189	0.232	0.265	0.273
CR-CNN	0.405	0.436	0.291	-	0.238	-
GSR-Net	0.387	0.613	0.285	0.175	0.283	0.349
SPAN	0.184	0.487	0.172	0.170	0.221	0.214
CAT-Net	0.276	0.352	0.134	0.102	0.138	0.200
MVSS-Net	0.452	0.638	0.453	0.260	0.292	0.419
Weakly-supervised methods.						
MIL-FCN	0.117	0.089	0.121	0.097	0.024	0.090
MIL-FCN+WSCL	0.172	0.270	0.178	0.193	0.110	0.185
Araslanov	0.112	0.102	0.127	0.094	0.026	0.092
Araslanov+WSCL	0.153	0.362	0.201	0.173	0.099	0.198
WSCL(re-trained)	0.150	0.305	0.169	0.062	0.015	0.140
Ours	**0.347**	**0.434**	**0.213**	**0.248**	**0.113**	**0.265**

Table 3: Compare against existing methods on pixel-level manipulation detection on F1 score (P-F1).

Pixel-level Localization Results. Tab. 3 showcases the comparative results for pixel-level manipulation localiza-

Method	Com-F1				
	CASIAv1	Columbia	COVER	IMD20	AVG
Unsupervised methods.					
NOI1	0.000	0.000	0.000	0.000	0.000
CFA1	0.000	0.000	0.000	0.000	0.000
Fully-supervised methods.					
H-LSTM	0.000	0.004	0.000	0.000	0.001
ManTra-Net	0.000	0.000	0.000	0.000	0.000
RRU-Net	0.023	0.000	0.000	0.000	0.006
CR-CNN	0.382	0.413	0.181	-	-
GSR-Net	0.042	0.042	0.000	0.026	0.028
SPAN	0.000	0.000	0.000	0.000	0.000
CAT-Net	0.345	0.406	0.149	0.144	0.261
MVSS-Net	0.566	0.711	0.317	0.300	0.474
Weakly-supervised methods.					
MIL-FCN	0.193	0.141	0.118	0.131	0.146
MIL-FCN+WSCL	0.280	0.386	0.268	0.252	0.296
Araslanov	0.194	0.140	0.133	0.046	0.125
Araslanov+WSCL	0.250	0.414	0.255	0.159	0.270
WSCL(re-trained)	0.242	0.457	0.258	0.111	0.267
Ours	**0.495**	**0.603**	**0.316**	**0.348**	**0.441**

Table 4: Overall performance on manipulation detection measured by combined F1 score (Com-F1), the harmonic mean of pixel-level F1 and image-level F1 on four test sets.

tion. Our approach outperforms weakly supervised MIL-FCN+WSCL algorithms, demonstrating a 43.2% superior average P-F1 metric. and the average performance on five datasets is comparable with the fully-supervised ManTra-Net and CAT-Net. Such a strong performance demonstrates the capability of our pixel-level manipulation localization.

Overall Performance. Tab. 4 provides overall detection and localization performance. Our method outperforms the weakly-supervised algorithm WSCL by a substantial margin of 45.9% and achieves the best overall performance (Com-F1) on the IMD20 dataset, which is close to real-world scenarios. Besides, our method surpasses the supervised algorithm, demonstrating that our method effectively maintains a balance between image-level and pixel-level image manipulation detection accuracy with only image-level labels.

Qualitative Results

As illustrated in Fig. 3, our method reveals that unsupervised methods produce noisy predictions, while both fully supervised and weakly supervised methods generate cleaner localization maps. For instance, the WSCL method predicts larger areas, encompassing the ground truth but is less precise compared to our M^{2}RL-Net method. Overall, our approach ensures clearer and more accurate localization.

Ablation Study

To evaluate the effectiveness of each component in our proposed approach, ablation experiments are conducted on the CASIAv1 and COVER datasets, with the results shown in Tab. 5. When RGB is replaced by a noise view, image-level detection scores (I-F1 and AUC) decrease, while pixel-level localization improves, and performance is further enhanced with the addition of multiview, indicating the benefit of noise cues in better artifact feature representation. The integration of an auxiliary classifier ($\mathcal{L}_{aux}$) aids in distinguishing

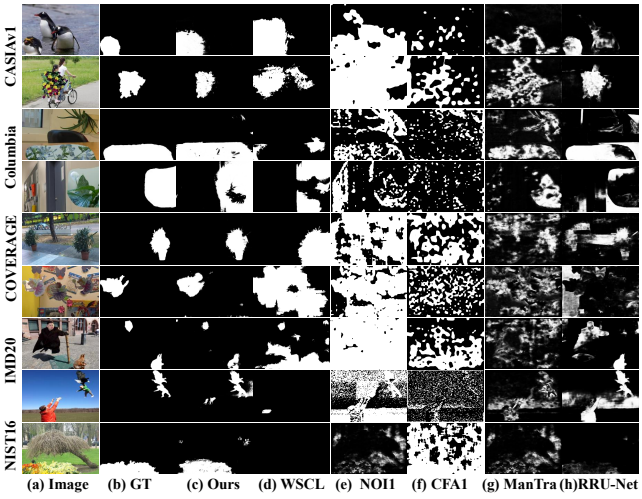

Figure 3: Qualitative comparison of our method (c) on five datasets with weakly-supervised (d), unsupervised (e)(f), and fully-supervised methods (g)(h).

Variant Models	CASIAv1			COVER		
	AUC	I-F1	P-F1	AUC	I-F1	P-F1
RGB View (baseline)	0.823	0.720	0.128	0.612	0.607	0.149
Noise View	0.842	0.634	0.151	0.622	0.609	0.164
MultiViews	0.916	0.813	0.280	0.675	0.562	0.156
+Aux. ($\mathcal{L}_{aux}$)	0.938	0.856	0.296	0.684	0.612	0.162
+PSL ($\mathcal{L}_{psl}$)	0.951	0.867	0.312	0.661	0.608	0.186
+FCL (w/ $\mathcal{L}_{self}$)	0.945	0.861	0.319	0.698	0.619	0.187
+FCL (w/ $\mathcal{L}_{multi}$)	0.943	0.860	0.323	0.701	0.636	0.197
+FCL ($\mathcal{L}_{fcl}$)	0.948	0.866	0.347	0.716	0.610	0.213

Table 5: Ablation results. Pixel-level F1 performance alongside Image-level F1 and AUC metrics are reported.

between real and forged images, enhancing detection performance, as evidenced by a 5.3% increase in I-F1 scores. Additionally, the inclusion of the PSL module significantly boosts overall performance, underscoring the module's effectiveness. Within the FCL framework, employing Self-contrast ($\mathcal{L}_{self}$) and Multiview-contrast ($\mathcal{L}_{multi}$) leads to the development of more discriminative features for image manipulation detection, enhancing localization results. The best performance is achieved when all components ($\mathcal{L}_{aux}$, $\mathcal{L}_{psl}$, and $\mathcal{L}_{fcl}$) are integrated, demonstrating the robustness and effectiveness of our comprehensive approach.

Conclusion

We introduce a novel approach for weakly supervised image forgery detection, dubbed as Multi-View and Multi-Level Relation Learning Network (M^{2}RL-Net). M^{2}RL-Net leverages patch-level self-consistency learning (PSL) and feature-level contrastive learning (FCL) to effectively use image-level annotations for precise forgery detection and localization. Our experimental results show that M^{2}RL-Net sets a new benchmark in the field, significantly improving detection accuracy and localization performance, while also paving the way for future advancements in image security.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (62273150, 62171323), Shanghai Natural Science Foundation (22ZR1421000), and Science and Technology Commission of Shanghai Municipality (22DZ2229004).

References

Araslanov, N.; and Roth, S. 2020. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4253–4262.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bappy, J. H.; Simons, C.; Nataraj, L.; Manjunath, B.; and Roy-Chowdhury, A. K. 2019. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE transactions on image processing*, 28(7): 3286–3300.

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.

Bi, X.; Wei, Y.; Xiao, B.; and Li, W. 2019. RRU-Net: The ringed residual U-Net for image splicing forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Chen, X.; Dong, C.; Ji, J.; Cao, J.; and Li, X. 2021. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14185–14193.

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.

Du, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2022. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4320–4329.

Ferrara, P.; Bianchi, T.; De Rosa, A.; and Piva, A. 2012. Image forgery localization via fine-grained analysis of CFA artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5): 1566–1577.

Franceschi, J.-Y.; Gartrell, M.; Dos Santos, L.; Issenhuth, T.; de Bézenac, E.; Chen, M.; and Rakotomamonjy, A. 2024. Unifying gans and score-based diffusion as generative particle models. *Advances in Neural Information Processing Systems*, 36.

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 (WACVW)*, 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.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

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, 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.

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.

Kwon, M.-J.; Yu, I.-J.; Nam, S.-H.; and Lee, H.-K. 2021. CAT-Net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 375–384.

Li, S.; Ma, W.; Guo, J.; Xu, S.; Li, B.; and Zhang, X. 2024. UnionFormer: Unified-Learning Transformer with Multi-View Representation for Image Manipulation Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12523–12533.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Lou, Z.; Cao, G.; Guo, K.; Zhu, H.; and Yu, L. 2024. Exploring Multi-view Pixel Contrast for General and Robust Image Forgery Localization. *arXiv preprint arXiv:2406.13565*.

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*.

Mahdian, B.; and Saic, S. 2009. Using noise inconsistencies for blind image forensics. *Image and vision computing*, 27(10): 1497–1503.

Niloy, F. F.; Bhaumik, K. K.; and Woo, S. S. 2023. CFL-Net: image forgery localization using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4642–4651.

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.

Pathak, D.; Shelhamer, E.; Long, J.; and Darrell, T. 2015. Fully Convolutional Multi-Class Multiple Instance Learning. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Qu, C.; Zhong, Y.; Liu, C.; Xu, G.; Peng, D.; Guo, F.; and Jin, L. 2024. Towards Modern Image Manipulation Localization: A Large-Scale Dataset and Novel Methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10781–10790.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ru, L.; Zhan, Y.; Yu, B.; and Du, B. 2022. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16846–16855.

Triaridis, K.; and Mezaris, V. 2024. Exploring multi-modal fusion for image manipulation detection and localization. In *International Conference on Multimedia Modeling*, 198–211. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Wu, Z.; Chen, J.; Han, X.; Shrivastava, A.; Lim, S.-N.; and Jiang, Y.-G. 2022. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2364–2373.

Wen, B.; Zhu, Y.; Subramanian, R.; Ng, T.-T.; Shen, X.; and Winkler, S. 2016. COVERAGE—A novel database for copy-move forgery detection. In *2016 IEEE international conference on image processing (ICIP)*, 161–165. IEEE.

Wu, H.; Chen, Y.; and Zhou, J. 2023. Rethinking image forgery detection via contrastive learning and unsupervised clustering. *arXiv preprint arXiv:2308.09307*.

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.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Yang, C.; Li, H.; Lin, F.; Jiang, B.; and Zhao, H. 2020. Constrained R-CNN: A general image manipulation detection model. In *2020 IEEE International conference on multimedia and expo (ICME)*, 1–6. IEEE.

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.

Zeng, K.; Cheng, R.; Tan, W.; and Yan, B. 2024. MGQFormer: Mask-Guided Query-Based Transformer for Image Manipulation Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6944–6952.

Zhai, Y.; Luan, T.; Doermann, D.; and Yuan, J. 2023. Towards generic image manipulation detection with weakly-supervised self-consistency learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22390–22400.

Zhang, L.; Xu, M.; Li, D.; Du, J.; and Wang, R. 2024a. CatmullRom Splines-Based Regression for Image Forgery Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7196–7204.

Zhang, X.; Li, R.; Yu, J.; Xu, Y.; Li, W.; and Zhang, J. 2024b. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11964–11974.

Zhang, Z.; Li, M.; and Chang, M.-C. 2024. A New Benchmark and Model for Challenging Image Manipulation Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7405–7413.

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.

Zhou, Y.; Wang, H.; Zeng, Q.; Zhang, R.; and Meng, S. 2024. Exploring weakly-supervised image manipulation localization with tampering Edge-based class activation map. *Expert Systems with Applications*, 249: 123501.

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.

Zhu, L.; Li, Y.; Fang, J.; Liu, Y.; Xin, H.; Liu, W.; and Wang, X. 2023. Weaktr: Exploring plain vision transformer for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2304.01184*.