Bridging Image Manipulation Localization and Generated Image Detection in Weakly-Supervised Setting

Lee YuSeop   Na YunJin   Noh HyunHo

1. Introduction

Recent advances in editing multimedia contents have emerged with the development of various deep learning based generative models, represented by Generative Adversarial Networks (GANs) [11] and Diffusion Models (DMs) [15, 32]. As AI technology is heavily leveraged to create and modify contents, attempts to develop the output of image generation and editing to appear more realistic have been highlighted. Recently, images generated by stable diffusion models such as DreamBooth[31] are so elaborate that they can be difficult to distinguish from authentic images. Especially, in the realm of AI-generated contents where the widespread availability of image editing techniques, ease of image manipulation has reached unprecedented levels. Image manipulation involves encompassing modification from simple alterations with copy-move, image splicing to advanced Deepfake generation facilitated by deceptive algorithms. However, this technical progress has also lead to critical challenge of the ability to discern authenticity from manipulated contents. With the potential for image forgery has grown exponentially, misinformation and fraudulent activities has escalated simultaneously.

While significant efforts to detect whether it is fake or not have been studied in the domain of generated image detection, show robustness across a variety of methodologies. However, the field of image manipulation localization lags behind in addressing real-world scenarios effectively. Even though existing approaches have achieved high accuracy from several benchmarks, it has been hampered by limitations inherent in existing datasets. These constraints include the necessity of pixel-level ground truth masks for fully supervised learning, absence of datasets with annotation synthesized through generative techniques and reliance on outdated benchmarks such as CASIA [8] proposed over a decade ago. Furthermore, the manipulation techniques in CASIA are limited to only two ways: copy-move and splicing. Consequently, existing researches would have encountered limitations in attempting to generalize and adapt to novel image benchmarks without ground truth masks [45], manipulated by generative models that aligns more closely with real-world scenarios.

In this context, our research endeavors to bridge these gap by proposing a novel approach aimed at achieving both robustness and generalization. Our method is designed to address two critical tasks concurrently: generated image identification and localized detection of manipulated regions within images, all without the constraint of labeled ground truth masks. By operating within a weakly-supervised framework, we expect to develop a versatile and generalized solution capable of robust performance across a wide range of manipulation scenarios. In contrast to conventional fully-supervised methods relying on pixel-level annotations, our proposed **weakly-supervised generated image manipulation detection** framework requires only binary image-level labels, aiming at significantly alleviating the burden of extensive manual annotation. Given the limitations that CNNs suffer from weak long-range and non-semantic feature modeling [24], we demonstrate our approach by leveraging both high-level semantic information from a Vision Transformer-based architecture [9] and low-level features from a ResNet-based model [14]. By fusing prediction maps generated by these models to create pseudo ground truth mask, we would take the advantages of both models' learning capabilities, thereby enriching our approach with both semantic and low-level features. In summary, through our approach, we would achieve the following main contributions:

1. Our work aims to address the challenging task of identifying whether images are either entirely or partially manipulated by generative models and localizing the manipulated regions within partially fake images. To the best of knowledge, it is the first attempt to generalize generated image detection in a weakly-supervised setting.
2. We propose a novel approach that combines multi-label discrepancy-aware knowledge distillation with adversarial learning. This allows our model to adapt to unseen data by effectively capturing characteristics of features, preserving robustness and generalization across different domains.
3. We integrate CNN-based and Transformer-based ap-

proaches for manipulation localization. By extracting activation maps from the CNNs and attention maps from the Transformer indicating patch correlations, it ensures effectively accurate learning even with limited labeled setting.

4. Through extensive experimental validation, we demonstrate that our framework achieves robust generalization in weakly-supervised localization of manipulated regions, even in the presence of unseen generative models by leveraging the proposed methods.

2. Related Work

2.1. Fake Image Detection

In the field of fake image detection, research has been conducted with diverse approaches according to different types of forgeries, broadly categorized into deepfake detection and generated image detection. [5] proposed a task for forged image detection in weakly-supervised setting, employing a CNN auto-encoder based approach with few-shot domain adaptation techniques. Addressing the challenges of cross-model generalization, [23, 36] investigated extracted discrepancies, focusing on artifacts such as high-frequency noise generated by CNN-based generators. [4, 25] explored more robust way to detect both generated by GANs and DMs, leveraging fingerprint and measuring feature distance between real and fake images, respectively. Further extending the research on diffusion models, [10, 37] enhanced generalization in challenging task, exploiting properties of the generative models with online augmentation, pixel prediction and reconstruction after inversion framework. Similarly, [6] also introduced cross-concept generalization task, calculating quality scores utilizing simple ResNet. Adversarial detection framework proposed in [44] aimed to generalize to unseen image generators by jointly performing teacher-student discrepancy-aware learning and generalized feature augmentation. Recently, [22] focused on detecting fake traces in generative models, learning trace representation with homogeneous and heterogeneous projection methods.

Furthermore, in the domain of deepfake detection, [39] presented knowledge distillation framework, leveraging latent space augmentation techniques and cross-domain augmentation. [12] developed a multi-stage detection framework, distinguishing authenticity of images and recognizing architectures between GANs and DMs. They utilized noise addition and reverse learning processes, demonstrating the superiority of ResNet over Vision Transformers.

While these papers contributed to the advancement of generalized fake image detection across different generative models and domain shifts, they have demonstrated that they predominantly rely on low-level features such as pixel prediction and typically engages in binary classification task. Therefore, simple models like ResNet achieved commendable performance. We observed that [25] was the only paper leveraging transformer models, which utilize high-level semantic information. We focused on the potential for further exploration with robustness and generalizability in integrating high-level features for more challenging detection task.

2.2. Image Manipulation Localization

Consequently, recent studies in image manipulation localization have attempted to address challenges through various approaches. [13] explored self-attention method for interaction modeling and feature fusion in CNNs. Based on this, [24] proposed the first attempt to leverage Vision Transformers for manipulation localization task. Subsequently, [43] introduced non-mutually contrastive learning techniques between feature maps of the manipulated regions with pre-training free encoder-decoder architectures. [1] addressed an innovative approach using encryption-based proactive scheme, exploiting separate detection and localization modules combined with transformer and CNNs. By incorporating a weakly-supervised setting that enables learning with only binary labels, [41] integrated multi-source and inter-patch consistency learning phase without intricate ground truth mask. [45] also demonstrated that effectiveness of patch-based approach robust on both fully-synthesized and partially manipulated images in weakly-supervised setting. Another contribution came from [35] rectified detection mechanism using different forensic filters with cues as input modalities, combining with encoder-decoder architectures through late and early fusion. In recent, [40] proposed a mask-guided query-based approach, utilizing a transformer decoder with query token to locate manipulations and [34] developed a method to learn forgery cues and manipulation maps without paired data. Their approach located and fused attention regions, focusing on locating exploitable cues.

Despite these studies highlighted efforts to enhance generalized manipulation localization with their own set of challenges, fundamental challenges still remain. Limitation arises from the benchmarks utilized for experiment and validation, as they were too small datasets and focused on basic manipulation techniques such as splicing. Additionally, existing methodologies solved the problem focusing on low-level discrepancies such as artifacts, thereby neglecting the nuances presented by modern generative models. Consequently, they may struggle to detect forgeries produced by these advanced techniques.

2.3. Weakly-Supervised Object Detection

We focused on how to demonstrate forgery detection in a weakly-supervised setting, exploiting attention and activation map-based object detection methods. This approach not only enhances detection capability but also provides interpretability by highlighting which parts of the model are focusing on. Previously, object detection primarily relied on Class Activation Maps (CAM), however, this approach presents several critical issues. Activation often occurs in the background areas or only activates part of the object, significantly hindering the learning. Therefore, recent studies in weakly-supervised object detection have addressed these challenges by considering various methods beyond limited resource settings. [21] presented a cooperative framework between detection and segmentation tasks, employing a collaboration loop with heatmaps and generative adversarial localization techniques. [2] introduced a method combining transformers and CNNs, leveraging the local perception capability while retaining global self-attention maps with cross-patch attention information. In a transformer-based approach, [30] focused on learning affinity from attention with pixel-adaptive refinement technique and [20]addressed the partial activation limitation inherent in CNN's local receptive fields, leveraging the attention weights from transformer to capture both low- and high-level spatial feature affinity. [38] explored a single-stage approach that emphasizes representation consistency between global and local views, integrating local semantics into transformer blocks.

Existing research has shown that there is a significant contrast between the advancements in generated image detection and the challenges encountered in image manipulation localization. The former has witnessed substantial progress, benefiting from active research and demonstrating its robustness across models and cross-domain, while the latter is hindered by the limitations inherent in available datasets and the rapid evolution of manipulation techniques. Thus, our work aims to bridge this gap by proposing methods to effectively handle both tasks within a weakly-supervised learning framework. By adopting principles from weakly-supervised object detection, we develop a framework that simultaneously address forgery classification and localization. Leveraging insights from existing researches, our approach emphasizes self-supervised learning, employing two distinct strategies: multi-label discrepancy-aware knowledge-distillation and localization with activation map and attention.

3. Method

3.1. Problem Definition and Architecture Overview

In this section, we introduce a novel framework aimed at enhancing robustness in the challenging task, detection and segmentation of generative manipulated images in a weakly-supervised setting, as illustrated in Figure 1. In the addressed task, there is no information without corresponding label whether a given image is real, partially manipulated, or fully generated. Then we divide the task into two primary challenges for efficiency, image identification and localization of manipulated regions within partially fake images. With the rapid emergence of generative methods, generators have advanced to precisely capture low-level features like pixel inconsistencies. Therefore, we endeavor to distinguish between real, partially fake, and completely fake images utilizing semantic information provided by transformers without degrading performance on unseen data. Especially in a weakly-supervised setting, it is crucial to utilize as much information as possible to accurately detect which region has been manipulated when dealing with partially fake images. To this end, we propose a method that combines a multi-label discrepancy-aware knowledge distillation approach with adversarial learning to easily adapt to unseen data by capturing its characteristics. For manipulation localization, we integrate CNN-based and transformer-based approach. We extract activation maps corresponding to the partially manipulated regions from the CNN and attention maps indicating correlations between patches from the transformer. The final result is derived by utilizing both of these maps. By integrating these methods, our framework aims to achieve robust generalization in weakly-supervised localization of manipulated regions, even in the presence of unseen generative models.

3.2. Multi-label Discrepancy-Aware Knowledge-Distillation

Rather than exploiting a single network to classify real, partially fake, and fully generated images, we propose a multi-label discrepancy-aware knowledge-distillation learning approach. This method enhances the detection of unseen data by controlling inconsistencies between the teacher network with general knowledge for classifying seen data and student networks, which learn specific characteristics about partially and completely generated images. For each image, we obtain image features $f = E(x)$ using a pre-trained feature extractor. These features f are then utilized in the subsequent training of the teacher, students, and generalizer networks. We denote the real image features as $f^R = E(x^R)$, the partially fake image features as $f^{PF} = E(x^{PF})$, and the fully fake image features as $f^{FF} = E(x^{FF})$, where x^R represents a real image, x^{PF} a partially fake image, and x^{FF} a fully fake image.

At first, we train the teacher network N_T on the training set using the cross-entropy loss function. The teacher network can classify the in-domain images as corresponding labels. After training N_T, we then freeze N_T and train two student networks, N_{S_1} and N_{S_2}, using three discrepancy loss functions under the guidance of the teacher network. The two student networks are designed to minimize discrepancies with the teacher network when processing real images and to maximize discrepancies when processing partially or fully fake images. By doing so, the student net-

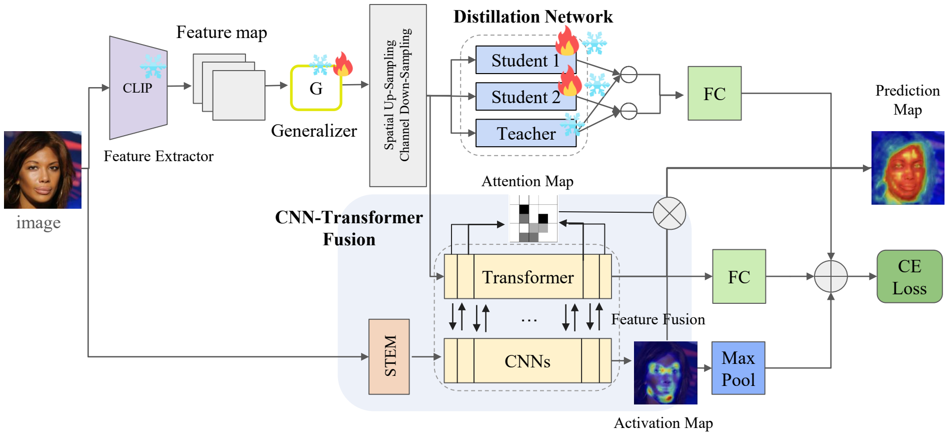


Figure 1. Comprehensive overview of our proposed framework. The architecture broadly consists of two parts: discrepancy-aware knowledge-distillation network and CNN-transformer fusion network.

works can learn both general information of real images and specific knowledge of partially and fully fake images.

For training with real images, we feed the real feature f^R into the teacher network to obtain the output $z^R_T = N_T(f^R)$, and we also input f^R into both student networks to obtain the outputs $z^R_{S_1} = N_{S_1}(f^R)$ and $z^R_{S_2} = N_{S_2}(f^R)$. When real images are provided as input, the discrepancy between N_T and $N_{S_{i=1,2}}$ should be minimized. The loss functions for real images are defined as:

$$\begin{aligned} L^R_{S_i}(z^R_T, z^R_{S_i}) &= \frac{1}{B}\sum_{b=1}^{B}(z^R_T - z^R_{S_i})^2 \\ &= \frac{1}{B}\sum_{b=1}^{B}(N_T(f^R) - N_{S_i}(f^R))^2 \end{aligned} \tag{1}$$

where i denotes the indices of student networks, and B denotes the batch size. We freeze the teacher network and train the student networks to minimize these losses for real image inputs.

When partially fake images are provided as input, the goal is to maximize the discrepancy between N_T and N_{S_2}, while minimizing the disparity between N_T and N_{S_1} to reflect their specificity towards the type of fake content. Therefore, the loss functions for partially fake images are modified to include a minimization term for N_{S_1}:

$$\begin{aligned} L^{PF}_{S_2}(z^{PF}_T, z^{PF}_{S_2}) &= \frac{1}{B}\sum_{b=1}^{B}[M - (z^{PF}_T - z^{PF}_{S_2})^2]_+ \\ = \frac{1}{B}\sum_{b=1}^{B}[M - (N_T(G^{PF}&(f^{PF})) - N_{S_2}(G^{PF}(f^{PF})))^2]_+ \end{aligned} \tag{2}$$

$$\begin{aligned} L^{PF}_{S_1}(z^{PF}_T, z^{PF}_{S_1}) &= \frac{1}{B}\sum_{b=1}^{B}(z^{PF}_T - z^{PF}_{S_1})^2 \\ &= \frac{1}{B}\sum_{b=1}^{B}(N_T(f^{PF}) - N_{S_1}(f^{PF}))^2 \end{aligned} \tag{3}$$

where $[.]_+ = \max(., 0)$, $z^{PF}_T = N_T(G^{PF}(f^{PF}))$, and $z^{PF}_{S_2} = N_{S_2}(G^{PF}(f^{PF}))$. We employ a regularization hyperparameter, denoted as margin M, which represents a desired minimum disparity that should be maintained between the augmented images and existing fake images for student models. If the gap exceeds M, the loss is reduced to zero, and learning no longer occurs. During training, features of manipulated images are augmented to enhance ability to discriminate and incorporate more plausible boundaries. This loss function aims to preserve significant output discrepancies between teacher and student networks when fake images are used as input.

Likewise, in the case of entirely fake images, the discrepancy between N_T and N_{S_1} should be maximized, while the discrepancy between N_T and N_{S_2} should be minimized:

$$\begin{aligned} L^{FF}_{S_1}(z^{FF}_T, z^{FF}_{S_1}) &= \frac{1}{B}\sum_{b=1}^{B}[M - (z^{FF}_T - z^{FF}_{S_1})^2]_+ \\ = \frac{1}{B}\sum_{b=1}^{B}[M - (N_T(G^{FF}&(f^{FF})) - N_{S_1}(G^{FF}(f^{FF})))^2]_+ \end{aligned} \tag{4}$$

$$
\begin{aligned}
L_{S_2}^{FF}(z_T^{FF}, z_{S_2}^{FF}) &= \frac{1}{B}\sum_{b=1}^{B}(z_T^{FF} - z_{S_2}^{FF})^2 \\
&= \frac{1}{B}\sum_{b=1}^{B}(N_T(f^{FF}) - N_{S_2}(f^{FF}))^2
\end{aligned}
\tag{5}
$$

where $z_T^{FF} = N_T(G^{FF}(f^{FF}))$, and $z_{S_2}^{FF} = N_{S_2}(G^{FF}(f^{FF}))$. By training the student networks with these loss functions, we ensure that they can effectively learn and generalize the differences between real, partial fake, and complete fake images. This method is designed so that the overall performance of the detection system is robust even when dealing with unseen generative models.

3.3. Adversarial Feature Generalizer.

Existing works [39, 44] have demonstrated that employing adversarial learning techniques can lead to better generalization across various tasks. Inspired by this, we integrate adversarial feature generalization into our framework to enhance the robustness of our model against unseen domains. Specifically, we propose two feature generalizers, G^{PF} for features of partial fake images and G^{FF} for features of fully fake images. Generalizer networks aims to capture and augment the generalization capabilities of the student models. By adversarially learning the extent to which fake features are available, this approach not only effectively adapts to unseen data but also strengthens its ability to identify between different types of manipulated images.

Training the Feature Generalizers. To train the adversarial feature generalizers, we modify the teacher N_T and student N_{S_1} and N_{S_2} networks. The objective is to minimize the discrepancy between the outputs of the teacher and each corresponding student network for fake images, thereby making it more challenging for the generalizers to augment features similar to those already seen.

Partially Fake Features. For the partial fake features, the training process involves using the discrepancy between the teacher and student outputs as adversarial guidance. When this discrepancy is large, it indicates that the generated features are similar to those already encountered by the existing generators in the training set. To encourage the generation of new, diverse features, we train G^{PF} to produce input features that minimize this discrepancy.

The loss function for training G^{PF} is defined as follows:

$$
\begin{aligned}
&L_{G^{PF}}^{PF}(z_T^{PF}, z_{S_1}^{PF}, z_{S_2}^{PF}) \\
&= \frac{1}{B}\sum_{b=1}^{B}([M - (z_T^{PF} - z_{S_1}^{PF})^2]_+ + (z_T^{PF} - z_{S_2}^{PF})^2) \\
&= \frac{1}{B}\sum_{b=1}^{B}([M - (N_T(G^{PF}(f^{PF})) - N_{S_1}(G^{PF}(f^{PF})))^2]_+ \\
&\qquad + (N_T(G^{PF}(f^{PF})) - N_{S_2}(G^{PF}(f^{PF})))^2)
\end{aligned}
\tag{6}
$$

Here, z_T^{PF}, $z_{S_1}^{PF}$, and $z_{S_2}^{PF}$ denote the outputs of the teacher and student networks when partially fake image features f^{PF} are inputted into the networks. By minimizing this loss, G^{PF} is trained to generate features that reduce the discrepancy between the teacher and student networks, thus making it difficult to maintain large output discrepancies when partially fake images are inputted.

Fully Fake Features. Similarly, for the fully fake features, we train G^{FF} with a loss function that minimizes the discrepancy between the teacher and student networks. The loss function for training G^{FF} is defined as follows:

$$
\begin{aligned}
&L_{G^{FF}}^{FF}(z_T^{FF}, z_{S_1}^{FF}, z_{S_2}^{FF}) \\
&= \frac{1}{B}\sum_{b=1}^{B}([M - (z_T^{FF} - z_{S_2}^{FF})^2]_+ + (z_T^{FF} - z_{S_1}^{FF})^2) \\
&= \frac{1}{B}\sum_{b=1}^{B}([M - (N_T(G^{FF}(f^{FF})) - N_{S_2}(G^{FF}(f^{FF})))^2]_+ \\
&\qquad + (N_T(G^{FF}(f^{FF})) - N_{S_1}(G^{FF}(f^{FF})))^2)
\end{aligned}
\tag{7}
$$

Here, z_T^{FF}, $z_{S_1}^{FF}$, and $z_{S_2}^{FF}$ denote the outputs of the teacher and student networks when fully fake image features f^{FF} are inputted into the networks. By minimizing this loss, G^{FF} is trained to generate features that reduce the discrepancy between the teacher and student networks for fully fake images.

Classification Process. After training the feature generalizers, the discrepancies between the teacher and student networks, derived from both real and fake images, are utilized to further enhance a CNN-Transformer stream classifier. This integrated classifier receives robustness information from the discrepancies, enriching its capabilities to discriminate between real, partially fake, and fully fake images with enhanced precision. During classification, the input images are first processed through the teacher and student networks to compute these discrepancies:

$$\hat{y} = N_C(concat((N_T(E(x)) - N_{S_1}(E(x)))^2, \\ (N_T(E(x)) - N_{S_2}(E(x)))^2)) \tag{8}$$

This classifier setup, combined with the initial discrepancy-based analysis, provides a comprehensive and robust framework for image manipulation detection. By employing these adversarial feature generalizers and integrating robustness information into the CNN-Transformer fusion networks, our framework achieves improved generalization and robustness, allowing it to effectively detect manipulated images even when dealing with unseen generative models.

3.4. CNN and Transformer Fusion

Convolutional Neural Networks (CNNs) excel at capturing local features effectively, while Transformer-based architectures are adept at extracting global semantic information across all patches through self-attention mechanisms. To leverage the strengths of both architectures, we propose a dual-stream framework that integrates CNN and Transformer branches for effective manipulation localization. We integrated the framework of existing work [20], based on Conformer architecture [26].

CNN Branch. The CNN branch is structured to efficiently capture local features through a series of convolutional operations. It comprises 12 layers, each consisting of a 1x1 down-projection convolution, a 3x3 spatial convolution, and a 1x1 up-projection convolution, connected via residual links. This hierarchical approach allows the model to incrementally gather detailed local features from the input images. The CNN branch is particularly effective at identifying local inconsistencies and fine-grained details in manipulated regions, such as texture anomalies and subtle pixel-level artifacts.

Transformer Branch. The transformer branch is designed to capture global representations using multi-head self-attention (MHSA) modules and multi-layer perceptrons (MLPs). Unlike conventional vision transformers, positional embeddings are omitted as the CNN branch already provides sufficient spatial information, allowing the Transformer to focus on global context. The transformer branch excels at identifying broader semantic inconsistencies, such as mismatched shadows or irregularities in object relationships, by leveraging its ability to capture long-range dependencies and global context.

Feature Fusion. To effectively combine the strengths of CNN and Transformer branches, we introduce a comprehensive feature fusion mechanism. This mechanism addresses dimensionality differences and facilitates continuous fusion of local and global features across multiple layers of the network. The fusion mechanism tackles the dimension differences between CNN feature maps $(C * H * W)$ and Transformer patch embeddings $((K + 1) * E)$. Downsampling using 1x1 convolutions and LayerNorm aligns the CNN features to the Transformer's dimension, while upsampling using 1x1 convolutions and BatchNorm aligns the Transformer embeddings back to the CNN's dimensions. This alignment process ensures effective fusion of local and global features. Beyond aligning feature dimensions, the fusion mechanism facilitates continuous interaction between local features from the CNN and global representations from the Transformer. This bidirectional flow ensures that both branches enhance each other's capabilities, resulting in a robust and comprehensive model. By combining detailed local cues and broader contextual information, our approach effectively captures the complexity of manipulated images. By integrating CNN and Transformer branches in a dual-stream framework with a robust feature fusion mechanism, our approach captures both local details and global context. This enhances the model's robustness and generalizability, making it well-suited for complex manipulation localization tasks in a weakly-supervised setting.

3.5. Localization

To effectively localize manipulated regions within images in a weakly-supervised setting, we combine the distinct strengths of CNNs and Transformers through an integrated approach leveraging both CNN activation maps and Transformer attention maps. This integration allows us to capitalize on the complementary capabilities of each architecture to enhance detection accuracy.

CNN Activation Maps. Class Activation Maps (CAMs) from the CNN branch represent discriminative image regions used by the CNN to identify specific classes. These maps highlight the most distinguishing features of an object but often suffer from the **partial activation** problem due to the CNN's local receptive field. This issue results in the emphasis of only the most salient parts of an object, potentially missing subtler, yet crucial, manipulated areas.

Transformer Attention Maps. In contrast, the Transformer branch generates attention maps that capture relationships and dependencies across all image patches, providing a broader view of global image context. These maps excel at identifying areas with semantic inconsistencies or abnormal interactions, which might not be as apparent in the local-focused CNN outputs.

Map Integration for Enhanced Localization. To address the limitations of partial activation in CNNs and to utilize the global sensitivity of Transformers, we merge the output from the CAM layer of the CNN with the attention maps from the Transformer. The CAM layer is configured to ensure that its output dimensions align with the number of patch tokens processed by the Transformer, facilitating an effective combination of both maps. The final segmentation is achieved by multiplying the CNN's activation maps with the Transformer's attention maps. This operation integrates the locally-focused, discriminative features highlighted by the CNN with the globally-aware, semantic correlations identified by the Transformer. By emphasizing areas that are suspicious in the CNN's view and corroborated by the Transformer's analysis, this method allows us to effectively pinpoint manipulated regions within the image. This integrated approach not only overcomes the inherent deficiencies of each individual model but also leverages their combined strengths to achieve a more accurate and robust localization of image manipulations in a weakly-supervised framework.

3.6. Inference

During the inference phase, our system effectively employs both the teacher-student and CNN-Transformer fusion networks to classify images. Importantly, the generalizers utilized for training the student's robustness are not included in the inference, ensuring a more streamlined and efficient process. Initially, the classifier determines whether an image is real, partially fake, or fully fake. If the image is classified as partially fake, the framework then proceeds to generate a prediction map specifically for the manipulated regions. Thus, after classifying an image's authenticity, our system uses the combined strengths of the CNN and Transformer models to produce a comprehensive prediction map.

4. Experiments

4.1. Experimental Settings

Datasets. In manipulated image detection tasks, CASIAv2.0 [8], CASIAv1.0 [7] are the most commonly used benchmarks. However, they consist of simple copy-move and splicing manipulations, fundamentally limiting their adaptability for comprehensive evaluation. Dolos [45], designed for generated manipulated localization, offers a diverse range of novel manipulation methods, including various generative models like LaMa [33], LDM [29], P2 [3] and Pluralistic [42] on large-scale CelebAHQ [17] and FFHQ [18] datasets, making it well-suited for our task. The proposed method in the paper also validates its performance in a weakly-supervised setting using this dataset. The detailed description of the datasets are shown in table 1. AutoSplice [16] serves as a novel resource for task of manipulated localization by generative models. This has not been extensively utilized in existing works, making it ideal for evaluating the robustness of different models through comparison. However, the data could not been evaluated since it was not approved by the authors. Therefore, we evaluated on different unseen generative methods of the dolos data. Furthermore, for generated fake image detection, we intend to utilize recently created datasets with large-scale data, ArtiFact [28], a fully synthesized datasets with generative models.

Table 1. Detailed dataset descriptions

Dataset	Authentic	Tampered	Copy-Move	Splicing
CASIAv2 [8]	7,491	5,063	3,235	1,828
CASIAv1 [7]	800	920	459	461
Dataset	**Authentic**	**Full Fake-DMs**	**Local Fake-DMs**	**Local Fake-Others**
Dolos [45]	20,700	20,000	85,300	21,600
ArtiFact [28]	964,989	1,531,749	-	-

Implementation Details. The architecture of the student, teacher, and generalizer networks is based on transformer block. Each transformer block consists of 4 layers and 384 embedding dimension, and 6 attention heads. The structure of transformer blocks in CNN-transformer fusion networks consist of 12 layers and 384 embedding dimension, and 6 attention heads. The learning rate for training is set to 5e-5, with a weight decay of 5e-4 and epsilon value of 1e-8. AdamW optimizer is exploited. For feature extraction, we utilize pre-trained CLIP:ViT [27]. To ensure robust generalization and fair comparison across tasks, we set the margin hyperparameter M to 4. This margin helps balance the range of losses among labels, preventing any single loss from becoming dominant. The training process rotates through five stages sequentially.

4.2. Comparative Analysis

4.2.1 Effectiveness on generalization

To evaluation the generalization capabilities of our proposed, we compared it against several baseline models across both in-domain and cross-domain settings. All experiments are conducted using the default settings to ensure consistency and fairness in evaluation.

Table 2. Evaluation results using the F1 score metric on in-domain and cross-domain of dolos dataset [45]. All methods are trained on Dolos-repaint-P2 dataset.

Dataset	Wang et al. [36]	TAR [19]	Ours
Dolos-Lama [45]	0.64	0.54	**0.68**
Dolos-LDM [45]	**0.83**	0.50	0.71
Dolos-Pluralistic [45]	**0.80**	0.53	0.69
Dolos-repaint-P2 [45]	**0.92**	0.82	0.81
Dolos-P2 [45]	0.68	0.41	**0.72**

In-domain and Cross-domain evaluation. Although our framework did not achieve the highest performance across all datasets, it exhibited a significantly smaller performance drop when applied to different datasets compared to other baselines. While the models by Wang et al. [36] and TAR [19] are designed for binary classification, our model is tailored for multi-label classification. Despite the increased complexity, our results remain highly competitive. Specifically, when trained solely on partially fake data, the

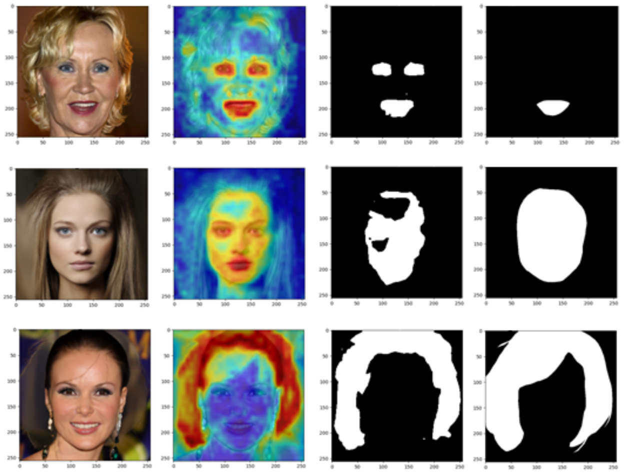


Figure 2. The leftmost column shows the input images, the second column from the left displays the prediction maps generated by our model, the third column from the left shows the binary prediction maps with a threshold of 0.5, and the rightmost column illustrates the ground truth segmentation maps.

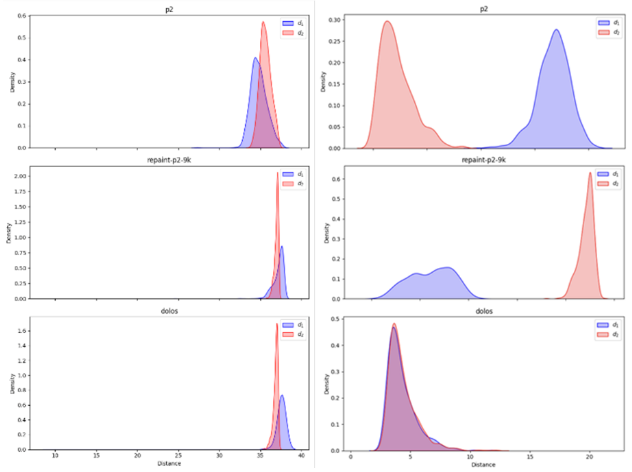


Figure 3. Distribution plot difference before and after training. Density distribution of full generated image at the top, partially manipulated image at the middle, and real data at the bottom.

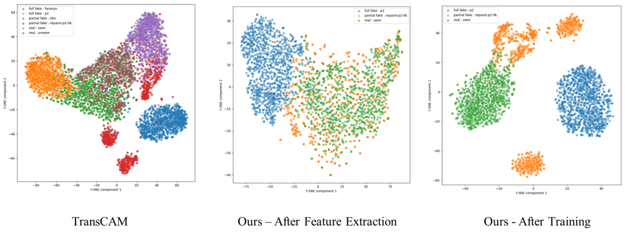
performance of other models significantly declined when tasked with detecting fully fake data. In contrast, our model maintained a relatively stable performance, demonstrating its robustness and effectiveness in diverse scenarios. This robustness and consistency across different types of fake data underscore the potential and reliability of our approach in real-world applications where the nature of the data may vary.

4.2.2 Visualization

Prediction map result. In Figure 2, the images in the leftmost column demonstrate how the input images are so intricately manipulated that it is difficult to identify the generated regions with the naked eye. Despite this, our model, which is trained solely on image-level labels, produces prediction maps (second column from the left) that closely resemble the ground truth (rightmost column). Furthermore, the binary prediction maps (third column from the left) with a threshold of 0.5 also show high accuracy. Our model does not arbitrarily predict facial or hair regions; instead, it focuses on identifying the genuinely manipulated areas. This is evident as the model demonstrates high accuracy in predicting the actual manipulated regions, even when the modifications are minimal and localized. This capability highlights the effectiveness of our model in detecting and segmenting manipulated parts, making it robust for various real-world scenarios.

Distribution plot. Figure 3 illustrates the impact of different types of image inputs on the distance between the output features of the student module and the teacher module during the initial and final stages of training in the knowledge distillation process. This is depicted through the distribution of the absolute differences in feature distances using Kernel Density Plots. From top to bottom, the rows labeled p2, repaint-p2-9k, and dolos correspond to scenarios where a full fake image, partial fake image, and real image are used as inputs, respectively. The columns d1 and d2 represent the feature distance differences between student 1 and the teacher, and student 2 and the teacher, respectively. During each stage of training, for the full fake image input scenario, student 1 was adversarially trained to diverge from the teacher, while student 2 was trained to converge towards the teacher. Conversely, for the partial fake image input scenario, student 1 was trained to converge towards the teacher, while student 2 was trained to diverge. Lastly, when a real image was used as input, both students were trained to mimic the teacher. Figure 3 demonstrates the intended outcomes, after training, compared to the initial stages where the feature distance differences between the students and the teacher were significant, the distribution of d1 for the full fake image scenario is positioned further along the x-axis than d2. For the partial fake image, the opposite is observed, and for the real image, both d1 and d2 are close to zero, indicating that the students successfully mimic the teacher.

T-SNE visualization. Figure 4 examines the practical impact of the knowledge distillation module on the CNN-transformer module at the feature level. The first plot is a TSNE representation of the feature distribution when the dolos data is trained on TransCAM. Technically, in this figure, it is not possible to make a direct comparison in an entirely identical environment since it includes different dataset [28], but it was referenced in the evaluation by in-



TransCAM

Ours – After Feature Extraction

Ours - After Training

Figure 4. T-SNE visualization result of TransCAM [20] trained with dolos[45], our model just after feature extraction with CLIP:ViT[27], final feature distribution of ours after training

cluding the same dataset. It shows that the features are well clustered, however, the boundaries between them are not clearly defined. The second plot shows the TSNE result after feature extraction using our method with a few epochs of training. As observed, the features of real and full fake images are well separated, but the features of real and partial fake images are not distinctly differentiated. This is because most partial fake images retain unmodified real image characteristics, leading to similar feature properties. In contrast, the final feature distribution plot in rightmost of Figure 4, after sufficient training, shows that the partial fake image features are also well separated from the real image features. This demonstrates that with adequate support of the knowledge distillation module, the image representation at the feature level continues to improve.

5. Conclusion

In this work, we propose a novel and intricate paradigm for detection and localization of regions within images manipulated by generated models with weakly-supervised settings on fully- and partially synthesized datasets [45]. We precisely designed to address the challenges of domain generalization and limited supervision in real-world scenarios, where manipulation detection proves especially challenging. Our framework combines a multi-label discrepancy-aware knowledge distillation approach with adversarial learning, enhancing the model's ability to adapt to unseen data. The integration of CNN-based and Transformer-based techniques for manipulation localization allows for precise detection of manipulated regions, even in weakly-supervised settings. By leveraging self-supervised learning mechanisms, we aim to enhance the generalization capabilities while reducing the reliance on extensive supervision of intricate annotation requirements. Through comprehensive experimentation, we demonstrate that our framework detects real, partially fake, and fully fake images robustly but also segment the manipulated region within images in various settings with only image-level labels. This makes our approach highly effective for real-world applications where sophisticated generative techniques continuously emerge. While our proposed framework leverages various techniques effectively, we will focus on further enhancing the adaptability and efficiency of our framework, exploring additional network architectures and more wide-ranging dataset for future work. Still, our work will mark a significant progress in the field of media forensics, paving the way for more adaptive and reliable solutions to protect authenticity while detecting contents manipulated with generative models in trustworthy digital media verification landscapes.

References

[1] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Malp: Manipulation localization using a proactive scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12343–12352, 2023. 2

[2] Zhiwei Chen, Changan Wang, Yabiao Wang, Guannan Jiang, Yunhang Shen, Ying Tai, Chengjie Wang, Wei Zhang, and Liujuan Cao. Lctr: On awakening the local continuity of transformer for weakly supervised object localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 410–418, 2022. 3

[3] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 7

[4] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[5] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 2

[6] Pantelis Dogoulis, Giorgos Kordopatis-Zilos, Ioannis Kompatsiaris, and Symeon Papadopoulos. Improving synthetically generated image detection in cross-concept settings. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, pages 28–35, 2023. 2

[7] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database 2010. 7

[8] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*, pages 422–426. IEEE, 2013. 1, 7

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 382–392, 2023. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[12] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Mastering deepfake detection: A cutting-edge approach to distinguish gan and diffusion-model images. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 2

[13] Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. Transforensics: image forgery localization with dense self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15055–15064, 2021. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[16] Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 893–903, 2023. 7

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 7

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 7

[19] Sangyup Lee, Shahroz Tariq, Junyaup Kim, and Simon S Woo. Tar: Generalized forensic framework to detect deepfakes using weakly supervised learning. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 351–366. Springer, 2021. 7

[20] Ruiwen Li, Zheda Mai, Zhibo Zhang, Jongseong Jang, and Scott Sanner. Transcam: Transformer attention-based cam refinement for weakly supervised semantic segmentation. *Journal of Visual Communication and Image Representation*, 92:103800, 2023. 3, 6, 9

[21] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly supervised object detection with segmentation collaboration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9735–9744, 2019. 3

[22] Ziyou Liang, Run Wang, Weifeng Liu, Yuyang Zhang, Wenyuan Yang, Lina Wang, and Xingkai Wang. Let real images be as a judger, spotting fake images synthesized with generative models. *arXiv preprint arXiv:2403.16513*, 2024. 2

[23] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021. 2

[24] Xiaochen Ma, Bo Du, Xianggen Liu, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023. 1, 2

[25] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 2

[26] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 367–376, 2021. 6

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7, 9

[28] Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2200–2204, 2023. 7, 8

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 7

[30] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. 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*, pages 16846–16855, 2022. 3

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[32] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[33] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 7

[34] Jiahe Tian, Peng Chen, Cai Yu, Xiaomeng Fu, Xi Wang, Jiao Dai, and Jizhong Han. Learning to discover forgery cues for

face forgery detection. *IEEE Transactions on Information Forensics and Security*, 2024. 2

[35] Konstantinos Triaridis and Vasileios Mezaris. Exploring multi-modal fusion for image manipulation detection and localization. In *International Conference on Multimedia Modeling*, pages 198–211. Springer, 2024. 2

[36] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 2, 7

[37] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 2

[38] Fangwen Wu, Jingxuan He, Yufei Yin, Yanbin Hao, Gang Huang, and Lechao Cheng. Masked collaborative contrast for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 862–871, 2024. 3

[39] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. *arXiv preprint arXiv:2311.11278*, 2023. 2, 5

[40] Kunlun Zeng, Ri Cheng, Weimin Tan, and Bo Yan. Mgqformer: Mask-guided query-based transformer for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6944–6952, 2024. 2

[41] Yuanhao Zhai, Tianyu Luan, David Doermann, and Junsong Yuan. Towards generic image manipulation detection with weakly-supervised self-consistency learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22390–22400, 2023. 2

[42] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 7

[43] Jizhe Zhou, Xiaochen Ma, Xia Du, Ahmed Y Alhammadi, and Wentao Feng. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22346–22356, 2023. 2

[44] Mingjian Zhu, Hanting Chen, Mouxiao Huang, Wei Li, Hailin Hu, Jie Hu, and Yunhe Wang. Gendet: Towards good generalizations for ai-generated image detection. *arXiv preprint arXiv:2312.08880*, 2023. 2, 5

[45] Dragoș-Constantin Țânțaru, Elisabeta Oneață, and Dan Oneață. Weakly-supervised deepfake localization in diffusion-generated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6258–6268, 2024. 1, 2, 7, 9