Beyond Fully Supervised Pixel Annotations: Scribble-Driven Weakly-Supervised Framework for Image Manipulation Localization

Songlin Li[1], Guofeng Yu[1], Zhiqing Guo[1,2*], Yunfeng Diao[3], Dan Ma[1], Gaobo Yang[4]

[1]School of Computer Science and Technology, Xinjiang University
[2]Xinjiang Multimodal Intelligent Processing and Information Security Engineering Technology Research Center
[3]School of Computer Science and Information Engineering, Hefei University of Technology
[4]College of Computer Science and Electronic Engineering, Hunan University

Abstract

Deep learning-based image manipulation localization (IML) methods have achieved remarkable performance in recent years, but typically rely on large-scale pixel-level annotated datasets. To address the challenge of acquiring high-quality annotations, some recent weakly supervised methods utilize image-level labels to segment manipulated regions. However, the performance is still limited due to insufficient supervision signals. In this study, we explore a form of weak supervision that improves the annotation efficiency and detection performance, namely scribble annotation supervision. We re-annotate mainstream IML datasets with scribble labels and propose the first scribble-based IML (Sc-IML) dataset. Additionally, we propose the first scribble-based weakly supervised IML framework. Specifically, we employ self-supervised training with a structural consistency loss to encourage the model to produce consistent predictions under multi-scale and augmented inputs. In addition, we propose a prior-aware feature modulation module (PFMM) that adaptively integrates prior information from both manipulated and authentic regions for dynamic feature adjustment, further enhancing feature discriminability and prediction consistency in complex scenes. We also propose a gated adaptive fusion module (GAFM) that utilizes gating mechanisms to regulate information flow during feature fusion, guiding the model toward emphasizing potential manipulated regions. Finally, we propose a confidence-aware entropy minimization loss. This loss dynamically regularizes predictions in weakly annotated or unlabeled regions based on model uncertainty, effectively suppressing unreliable predictions. Experimental results show that our method outperforms existing fully supervised approaches in terms of average performance both in-distribution and out-of-distribution.

Code — https://github.com/vpsg-research/SCAF

Introduction

Maliciously manipulated images can spread false information and seriously threaten social harmony. To safeguard information security and uncover the truth behind image manipulation, image manipulation localization (IML) technology, which focuses on accurately segmenting the manipulated regions, has been extensively studied.

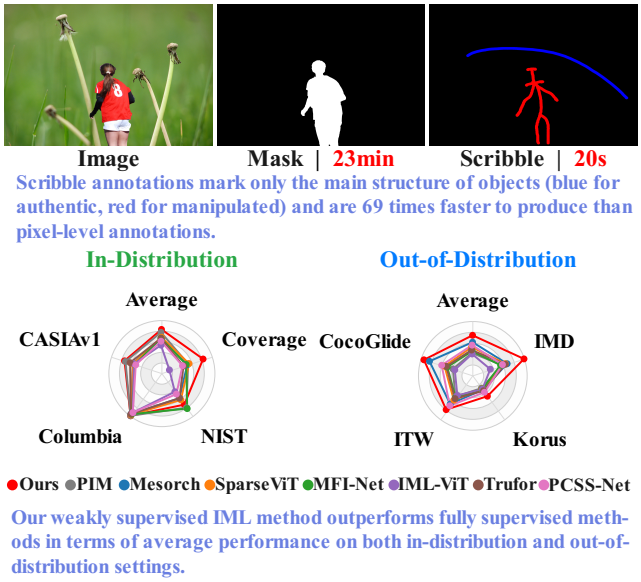

Figure 1: Comparative results between our model and existing IML methods in labeling efficiency and performance.

In recent years, deep learning techniques have significantly advanced the performance of IML methods, particularly when supported by large-scale pixel-level annotated datasets. Fully supervised methods can leverage dense pixel annotations to learn fine-grained manipulation features, achieving state-of-the-art localization results on various benchmark datasets. However, acquiring high-quality pixel-level annotations is both time-consuming and labor-intensive, which greatly limits the scalability and practical application of these methods. In real-world scenarios, obtaining large-scale high-precision pixel annotations is often infeasible, especially as data diversity and scale continue to expand. To reduce annotation costs, weakly supervised IML methods have emerged in recent years, typically relying on image-level labels to guide the localization of manipulated regions. While such methods alleviate the dependence on dense annotations, their limited supervision signals lead to clear bottlenecks in localization accuracy and generalization capability. A substantial performance gap still exists be-

*Corresponding author: Zhiqing Guo, guozhiqing@xju.edu.cn

tween weakly supervised and fully supervised methods.

To address these challenges, this paper explores a form of weak supervision, namely scribble annotations, which takes into account the efficiency of labeling and the amount of supervision information, as shown in Fig. 1. Scribble annotations enable rapid labeling of large-scale IML datasets while providing informative cues for localization tasks. We re-annotated several mainstream IML datasets, including 5,123 images from CASIAv2 (Dong, Wang, and Tan 2013), 70 images from Coverage (Wen et al. 2016), 130 images from Columbia (Hsu and Chang 2006), and 414 images from NIST16 (Guan et al. 2019), resulting in a total of 5,737 images. This constitutes the first scribble-based IML (Sc-IML) dataset. During annotation, annotators used CVAT to draw scribbles on the manipulated regions based on their first impression, without referring to the ground truth. To ensure high-quality data, the annotation process was cross-verified by three reviewers. Each image took approximately 20 seconds to annotate. However, existing studies have not reported the time required to manually annotate a pixel-level mask for a manipulated image. To address this issue, we organized 10 experienced computer vision researchers. Each researcher randomly selected 10 images from the training dataset for pixel-level annotation, with an average annotation time of approximately 23 minutes per image. This shows that scribble annotation is 69 times faster than pixel-level annotation. Although scribble annotations are significantly more efficient and convenient compared to pixel-level masks, they still face two major limitations: (1) Scribble annotation is a highly subjective form of weak supervision. Different annotators may have varying interpretations of the manipulated regions, boundaries, and details, resulting in significant inconsistency among the scribble annotations. (2) Scribble labels offer limited pixel-level supervision, causing the model to lack confidence in classifying unmarked regions and leading to prediction uncertainty.

Based on this, we propose the first weakly supervised IML framework utilizing scribble annotations, namely SCAF. To address the challenges of prediction inconsistency and uncertainty under scribble-based weak supervision in IML datasets, we introduce a series of innovative strategies. To resolve inconsistency, we employ self-supervised training with a structural consistency loss, which constrains the model to produce consistent predictions under multi-scale and augmented inputs. We also propose a prior-aware feature modulation module (PFMM) that adaptively integrates prior information from both manipulated and authentic regions for dynamic feature adjustment, and incorporate coordinate attention to efficiently model spatial dependencies, thereby significantly enhancing the discriminability and scene adaptability of feature representations. To tackle prediction uncertainty, we propose a gated adaptive fusion module (GAFM) that regulates information flow through multi-branch channel splitting and adaptive dynamic fusion. This module not only highlights critical features and suppresses redundancy, but also leverages scribble supervision to guide the model's attention toward potential manipulated regions. Additionally, we propose a confidence-aware entropy minimization loss ($\mathcal{L}_{CEM}$), which dynamically filters model outputs based on uncertainty and applies adaptive entropy regularization to weakly annotated or unlabeled regions. This effectively suppresses unreliable predictions and improves model confidence and generalization in key areas. In summary, our contributions to IML are as follows:

- We propose the Sc-IML, the first scribble annotated dataset specifically designed for weakly supervised IML. Sc-IML effectively bridges the gap between costly pixel-level annotations and coarse image-level supervision by providing valuable spatial cues for the development and evaluation of weakly supervised IML methods, thereby advancing research in this field.
- We propose the first weakly supervised IML framework based on scribble annotations, which outperforms existing fully supervised methods in terms of both in-distribution and out-of-distribution average performance.
- We propose a PFMM that adaptively integrates prior information from both manipulated and authentic regions to achieve dynamic feature modulation. We further incorporate coordinate attention to efficiently model spatial dependencies, thereby significantly enhancing feature discrimination and scene adaptability.
- We propose a GAFM that regulates information flow through multi-branch channel splitting and adaptive dynamic fusion. This module not only highlights critical features and suppresses redundancy, but also leverages scribble annotations to guide the model's attention toward potential manipulated regions.
- We propose a confidence-aware entropy minimization loss, which adaptively regularizes predictions in weakly annotated or unlabeled regions, significantly suppresses unreliable predictions, and enhances model confidence and generalization in critical areas.

Related Work

Fully supervised image manipulation localization: With the continuous development of deep learning and the emergence of large-scale datasets, fully supervised IML methods have achieved remarkable results. For example, (Guillaro et al. 2023) proposed an IML framework that combines RGB data with noise-resistant fingerprints for IML. (Chen et al. 2024) effectively enhances the detection of boundary artifacts in IML by refining edge features and leveraging multi-feature fusion. (Gu et al. 2024) employs a Siamese network to extract illumination features and integrate them with a U-shaped network to localize manipulated regions. (He et al. 2024) embeds manipulation cues into similarity features to guide the matching process toward suspicious regions. (Kong et al. 2025a) exploits pixel inconsistencies to model global and local dependencies, enabling robust generalization to diverse manipulations.

Weakly supervised image manipulation localization: To reduce annotation costs, weakly supervised IML methods have been developed to segment manipulated regions using only image-level labels during training. (Zhai et al. 2023) applies self-consistency learning with multi-source cues to improve localization. (Zhou et al. 2024b) iteratively

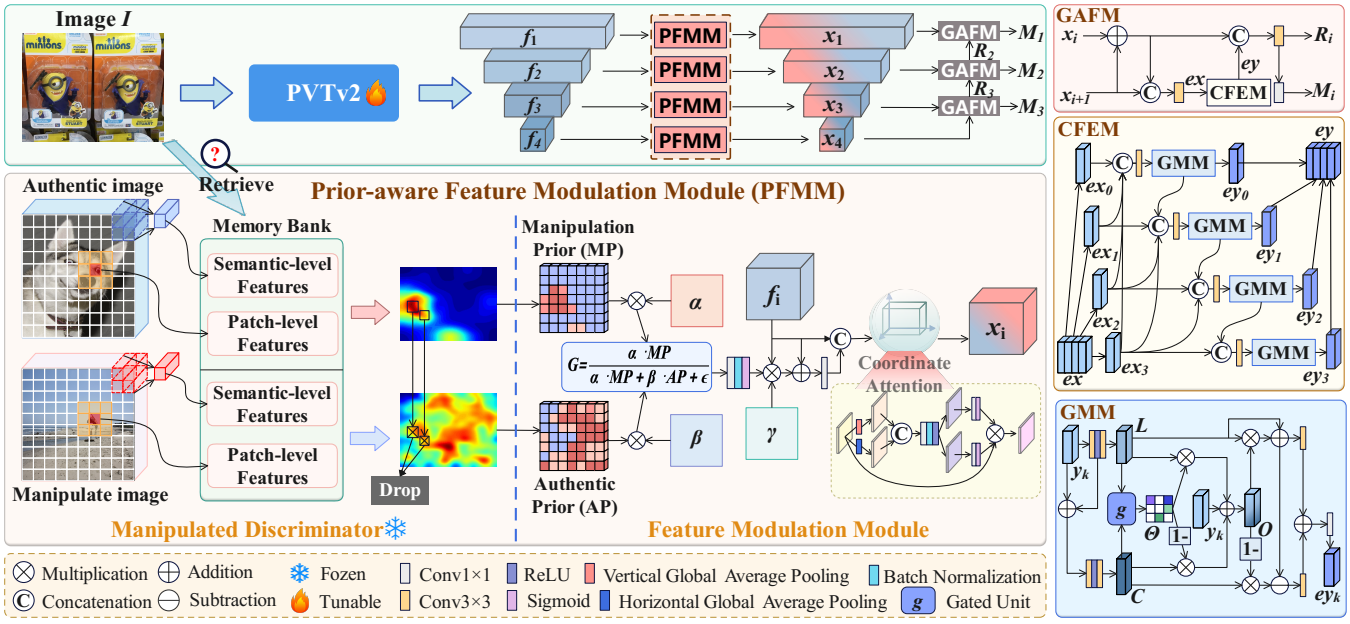

Figure 2: The overall architecture of the proposed SCAF. The model comprises two key modules: the prior-aware feature modulation module (PFMM) and the gated adaptive fusion module (GAFM). It is worth noting that the PFMM consists of a manipulated discriminator (MD) and a feature modulation module (FMM).

refines pseudolabels to sharpen boundaries. (Zhu, Li, and Wen 2025) introduces a cross-contrastive multi-stream fusion network that dispenses with pixel-level annotations. (Bai 2025) proposes WSCCL, which exploits multiple point correspondences for IML.

Although fully supervised methods achieve high localization accuracy, they rely on large-scale high-quality pixel-level annotations, which are costly and time-consuming to obtain. Their scalability is further limited by the lack of such data in real-world scenarios. Weakly supervised IML methods using image-level labels greatly reduce annotation costs but lack spatial cues, making precise localization difficult and often missing subtle manipulations. To address these challenges, we propose a novel weakly supervised IML approach with scribble annotations. Scribbles provide crucial spatial information to the model, guiding it to learn the spatial distribution of manipulated regions while significantly improving annotation efficiency. Our model outperforms existing fully supervised methods in terms of average performance, both in-distribution and out-of-distribution.

Methodology

Overview

The overall architecture of SCAF is shown in Fig. 2. Specifically, the input image $\boldsymbol{I}$ is first processed by PVTv2 (Wang et al. 2022) to extract multi-scale features $\boldsymbol{f}_i$, and these features are then modulated in the prior-aware feature modulation module (PFMM) using the prior to integrate them into $\boldsymbol{x}_i$. Finally, the features are fused by the gated adaptive fusion module (GAFM) to produce the final predicted mask. The training process of the model is illustrated in

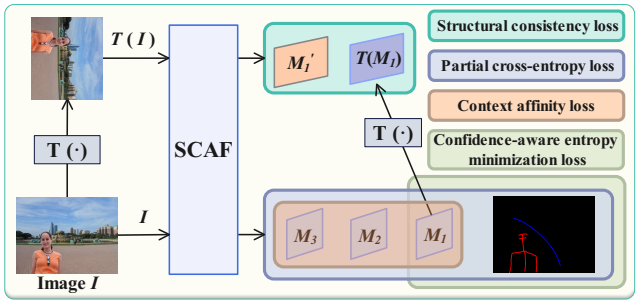

Figure 3: The training process of the proposed model. $\boldsymbol{T}(\cdot)$ denotes random rotations, scaling, and flipping.

Fig. 3. An image $\boldsymbol{I}$ is input into the SCAF to generate the primary output $\boldsymbol{M}_1$ as well as auxiliary outputs $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$. Additionally, a randomly transformed version of $\boldsymbol{I}$, denoted as $T(\boldsymbol{I})$, which may include rotation, scaling, or flipping, is also fed into the SCAF to obtain $\boldsymbol{M}_1'$. This, together with $\boldsymbol{M}_1$, is used to compute the structural consistency loss ($\mathcal{L}_{SC}$) (He et al. 2023). Furthermore, $\boldsymbol{M}_1$ is jointly supervised by the partial cross-entropy loss ($\mathcal{L}_{PCE}$), context affinity loss ($\mathcal{L}_{CA}$) (Obukhov et al. 2019), and confidence-aware entropy minimization loss ($\mathcal{L}_{CEM}$), while $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$ are optimized using only $\mathcal{L}_{CA}$ and $\mathcal{L}_{PCE}$.

Prior-aware Feature Modulation Module

For scribble annotations, annotators typically mark the manipulated regions in an image by freely drawing rough scribbles. Since this annotation method heavily relies on personal subjective judgment, different annotators may have varying

understandings of the manipulated regions, interpretations of boundaries, and levels of attention to detail. Even for the same image, the content of the scribble annotations can differ significantly. This subjectivity and randomness lead to considerable label inconsistency within the training set, which in turn affects the model's ability to accurately identify manipulated regions and generalize to new data. In complex scenarios, this often results in unstable and unreliable predictions. To address the inconsistency caused by subjective scribble annotations, we propose a prior-aware feature modulation module (PFMM). As shown in the lower left of Fig. 2, PFMM is composed of two parts: the manipulated discriminator (MD) and the feature modulation module (FMM). The core idea is to utilize manipulated priors (MP) and authentic prior (AP) generated by the MD, and then leverage these priors within the FMM to dynamically modulate the features $\boldsymbol{f}_i$ in a targeted manner. By explicitly introducing prior knowledge, PFMM can effectively mitigate the noise and bias introduced by subjective scribble annotations, guiding the model to focus on more objective region distributions. This mechanism enhances the model's discriminability for manipulated regions and improves prediction consistency and robustness.

1) Manipulated discriminator (MD): Existing research (Roth et al. 2022) has demonstrated that patch-level features are effective for detecting subtle anomalies. However, they lack explicit suppression of semantic bias, which becomes particularly evident in cross-domain scenarios. To address this limitation, we propose a selective suppression mechanism based on semantic-level features, which builds upon patch-level representations. By memorizing multi-scale semantics from both authentic and manipulated images, our method adaptively suppresses features that are highly similar to those stored in the memory bank during inference. This enables the generation of precise authentic region priors and manipulated region priors. Specifically, taking the training process of authentic images as an example, we first extract multi-layer features using the PVTv2. Each layer's feature map is then partitioned into overlapping patches with a stride of 1, resulting in a set of local patch features $\boldsymbol{p}_j^s$. To further incorporate local contextual information, we perform weighted fusion of each patch and its eight spatial neighbors to obtain $\tilde{\boldsymbol{p}}_j$, as formulated below:

$$\tilde{\boldsymbol{p}}_j^s = \sum_{n \in \mathcal{N}(j)} w_n \boldsymbol{p}_n^s \tag{1}$$

where $\mathcal{N}(j)$ denotes the 3×3 neighborhood centered at j, and w_j represents the weighting coefficient. The collection of $\tilde{\boldsymbol{p}}_j^s$ constitutes the patch-level feature $\boldsymbol{P}_s$ for an image. We store the patch-level features $\boldsymbol{P}_s$ of all images in the training set to construct a patch-level memory bank $\mathcal{B}_p$:

$$\begin{cases} \boldsymbol{P}_s = \left\{ \tilde{\boldsymbol{p}}_j^s \mid j = 1, \ldots, T \right\} \\ \mathcal{B}_p = \left\{ \boldsymbol{P}_s \mid s = 1, \ldots, N \right\} \end{cases} \tag{2}$$

where T denotes the total number of local patch features in a sample. N denotes the number of training samples. For the construction of the semantic-level memory bank $\mathcal{B}_s$, we first extract features from the authentic training samples using the backbone network, and reduce the channel dimension of the extracted features to 256 to obtain the feature set $\{\boldsymbol{v}_m\}_{m=1}^N$. These features are then subjected to L2 normalization. Finally, the semantic features of all training images are stored to form the memory bank:

$$\begin{cases} \tilde{\boldsymbol{v}}_m = \boldsymbol{v}_m \,/\, \|\boldsymbol{v}_m\|_2 \\ \mathcal{B}_s = \{\tilde{\boldsymbol{v}}_m \mid m = 1, \ldots, N\} \end{cases} \tag{3}$$

The model constructs a memory bank of authentic region features by learning from a large number of real images. However, this bank is filled with highly similar features, resulting in redundancy and bias, which leads the model to overgeneralize dominant authentic region features. During inference, this mechanism causes subtle manipulated features to be easily overwhelmed by the abundance of highly matching authentic features, thereby reducing the saliency of manipulated regions and ultimately lowering localization accuracy. To address this issue, we actively identify and suppress redundant authentic feature noise, which relatively amplifies and highlights manipulated features that represent inconsistencies. In this way, the model shifts from passively learning authentic image features to actively capturing the distinctive traces of manipulation. The details are as follows:

$$\boldsymbol{q}_{sup} = \left(1 - \max_m \left(\frac{\boldsymbol{q} \times \boldsymbol{v}_m}{\|\boldsymbol{q}\|_2 \, \|\boldsymbol{v}_m\|_2} \right) \right) \times \boldsymbol{q} \tag{4}$$

where $\boldsymbol{q}$ denotes the current input feature, which is compared with each feature $\boldsymbol{v}_m$ in $\mathcal{B}_s$ to compute the maximum cosine similarity. This process suppresses regions that are overly similar to authentic features from the training set and highlights manipulated features, resulting in $\boldsymbol{q}_{sup}$. Finally, the Euclidean distance between $\boldsymbol{q}_{sup}$ and its nearest feature $\boldsymbol{P}_s$ in $\mathcal{B}_p$ is calculated. It can be formulated as follows:

$$\mathrm{Score}(\boldsymbol{q}_{sup}) = \|\boldsymbol{q}_{sup} - \boldsymbol{P}_s\|_2 \tag{5}$$

By scoring each location in the input image, we obtain a prior map of manipulated regions. Similarly, training on manipulated images yields a prior map for authentic regions. However, since manipulated images contain both authentic and manipulated areas, the authentic prior may include false activations. To address this, we compute the cosine similarity between the manipulated and authentic priors, removing falsely activated regions to obtain a purified authentic prior.

2) Feature modulation module (FMM): FFM is designed to adaptively modulate the feature $\boldsymbol{f}_i$ using prior information, thereby alleviating the adverse effects caused by the inconsistency of scribble annotations. Specifically, FFM utilizes learnable manipulated region weight α and authentic region weight β to perform weighted normalization of the prior knowledge, resulting in the computation of the manipulated region probability response $\boldsymbol{G}$, as formulated below:

$$\boldsymbol{G} = \frac{\alpha \times MP}{\alpha \times MP + \beta \times AP + \epsilon} \tag{6}$$

where ϵ is a small constant added to prevent division by zero. This controllable probabilistic modeling approach allows the module to adaptively adjust its sensitivity to different regions based on the data distribution, effectively reducing the impact of label inconsistency or prior noise on

performance. Then, $\boldsymbol{G}$ is passed through a 1×1 convolution, batch normalization, and a sigmoid activation function to generate an enhanced feature map $\boldsymbol{Ge}$ for each spatial location. This is further regularized residually by a learnable parameter γ to obtain the feature $\boldsymbol{F}$, as follows:

$$\boldsymbol{F} = \mathrm{Conv1}(\boldsymbol{f_i} + \gamma \times \boldsymbol{Ge} \times \boldsymbol{f_i}) \tag{7}$$

where $\mathrm{Conv1}$ denotes a 1×1 convolution. Finally, coordinate attention is applied to the concatenated features of $\boldsymbol{F}$ and $\boldsymbol{f}_i$, resulting in the output feature $\boldsymbol{x}_i$. Coordinate attention captures both spatial location information and channel interdependencies, further enhancing the model's ability to localize and discriminate manipulated regions.

In summary, PFMM explicitly incorporates manipulated and authentic priors to effectively alleviate the adverse effects of inconsistency and label noise introduced by scribble annotations. By leveraging prior-guided feature modulation, the model is able to focus on more objective and reliable regional cues, thereby enhancing both the accuracy and generalization ability of tampering localization.

Gated Adaptive Fusion Module

In scribble-based weakly supervised IML, annotations are sparse and irregular, making it challenging for models to fully localize manipulated regions. This incomplete supervision introduces significant uncertainty during training and hampers the detection of subtle manipulations, often leading to missed or false detections. Furthermore, manipulated regions can exhibit diverse spatial distributions and weak visual clues, which are hard to capture using conventional feature aggregation or single-scale modeling. To address these issues, we propose a gated adaptive fusion module (GAFM), as shown on the right side of Fig. 2, which consists of a core channel-aware feature enhancement module (CFEM) and a gated modulation module (GMM). The GAFM aggregates multi-scale contextual features and employs both the CFEM and GMM to group channels and progressively perform gated fusion, thereby adaptively suppressing feature uncertainty and enhancing the richness and discriminability of the feature representations. Specifically, the fused feature $\boldsymbol{e}_x$, obtained by merging $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+1}$, is first input into CFEM, where it is evenly split into four groups along the channel dimension.

$$\boldsymbol{ex} = [\boldsymbol{ex}_0, \boldsymbol{ex}_1, \boldsymbol{ex}_2, \boldsymbol{ex}_3], \quad \boldsymbol{ex}_k \in \mathbb{R}^{B \times \frac{C}{4} \times H \times W} \tag{8}$$

Then, each group feature $\boldsymbol{ex}_j$ is concatenated with its neighboring groups or the output of the previous GMM, and sequentially passed through a 3×3 convolution and the GMM, forming a progressive information fusion flow.

$$\begin{cases} \boldsymbol{y}_k = \mathrm{Conv3}(Cat(Cat\mid_{z=k}^{3} \boldsymbol{ex}_z, \mathbb{I}(k > 0) \times \boldsymbol{ey}_{k-1})) \\ \boldsymbol{ey}_k = \mathrm{GMM}\,(\boldsymbol{y}_k) \end{cases} \tag{9}$$

where $\mathrm{Conv3}$ denotes a 3×3 convolution. $Cat\mid_{z=k}^{3}$ denotes the concatenation of all $\boldsymbol{ex}_n$ for $n = j$ to 3 in sequence. $\mathbb{I}(k > 0)$ denotes the indicator function, which returns 1 if $k > 0$, and 0 otherwise. The GMM employs a gating mechanism to achieve adaptive recalibration and interaction between local and global features, thereby enhancing the modeling capability for fine-grained representations. Specifically, the feature $\boldsymbol{y}_k$ is first processed through two separate sequences of 3×3 convolutional layers followed by ReLU activations, resulting in two distinct feature branches.

$$\begin{cases} \boldsymbol{L} = \mathrm{Conv3}(ReLU(\mathrm{Conv3}(\boldsymbol{y}_k))) \\ \boldsymbol{C} = \mathrm{Conv3}(ReLU(\mathrm{Conv3}(\boldsymbol{L} + \boldsymbol{y}_k))) \end{cases} \tag{10}$$

Subsequently, the extracted local-detail feature $\boldsymbol{L}$ and contextual semantic feature $\boldsymbol{C}$ are concatenated along the channel dimension and fed into a gating unit $g(\cdot)$ (Zhu et al. 2023). By effectively leveraging the complementary relationship between these two types of features, the gating unit adaptively generates a more accurate gating coefficient α, enabling precise modulation and enhancement of features of the manipulated region. It can be formulated as:

$$\begin{cases} \theta = g(Cat(\boldsymbol{L},\, \boldsymbol{C})) \\ \boldsymbol{O} = \boldsymbol{y}_j + \theta \times \boldsymbol{L} + (1 - \theta) \times \boldsymbol{C} \end{cases} \tag{11}$$

Due to the limited pixel-level supervision provided by scribble-based annotations, the model may confuse features of manipulated regions with those of the authentic background, leading to ambiguous localization. To address this, we construct a complementary reverse mask $(1 - \boldsymbol{O})$ from the feature map $\boldsymbol{O}$ and combine it with feature $\boldsymbol{C}$ to obtain the authentic background features. We enhance the manipulated region's feature representation $\boldsymbol{O}$ by combining it with feature $\boldsymbol{L}$ and applying a residual connection. Subsequently, the difference between the manipulated-enhanced feature $\boldsymbol{Me}$ and the authentic background features is computed to highlight subtle or latent anomalies within the manipulated regions, resulting in the differential feature $\boldsymbol{D}$. Then, the manipulated-enhanced feature $\boldsymbol{Me}$ is fused with the differential feature $\boldsymbol{D}$ to obtain the output feature $\boldsymbol{ey}_j$, which can be expressed as $\boldsymbol{ey}_k = \mathrm{Conv1}(\boldsymbol{Me} + \boldsymbol{D})$.

The GMM enables the model to adaptively focus on manipulated regions, enhancing manipulated features while suppressing background features. Finally, we concatenate the outputs from multiple stages of the GMM to obtain the feature $\boldsymbol{e}y$, which can be expressed as $\boldsymbol{ey} = Cat\mid_{k=0}^{3} \boldsymbol{ey}_k$.

Since each stage of the GMM focuses on adaptive modeling of specific sub-channel features, these features are inherently complementary. By concatenating them along the channel dimension, the model maximally preserves the heterogeneous information across groups, thereby providing richer contextual cues for subsequent integration and enhancing the perception and representation of complex manipulated regions.

Confidence-aware Entropy Minimization Loss

In scribble-based weakly supervised IML, the sparsity of supervision signals makes it difficult to adequately constrain the model's learning. While entropy minimization is an effective strategy for leveraging unlabeled regions, blindly applying it may cause the model to become "over-confident" on unreliable predictions, amplifying errors and resulting in unstable training. To address this issue, we propose a confidence-aware entropy minimization loss ($\mathcal{L}_{CEM}$). This loss applies entropy minimization only to unlabeled pixels whose current predictions are deemed reliable based on

Method	CASIAv1	Coverage	NIST	IMD	Columbia
WSCL	0.153	0.201	0.099	0.173	0.362
EdgeCAM	0.301	0.262	0.254	0.242	0.470
SOWCL	0.334	0.239	0.288	0.259	0.385
WSCCL	0.349	0.281	0.278	0.259	0.516
Ours	**0.716**	**0.827**	**0.721**	**0.580**	**0.979**

Table 1: Comparison with other advanced weakly supervised IML methods.

a confidence threshold. Additionally, a weak regularization term is applied to the annotated regions to directly encourage the model to produce predictions with higher confidence and sharper boundaries. Specifically, the model's primary output is $\boldsymbol{M}_1 \in \mathbb{R}^{B\times 1\times H\times W}$, where each pixel is associated with a probability value $m_t \in [0, 1]$. The Shannon entropy $\mathcal{H}(\cdot)$ for each pixel is defined as follows:

$$\mathcal{H}\left(\boldsymbol{m}_t\right) = -\left[\boldsymbol{m}_t \log \boldsymbol{m}_t + (1-\boldsymbol{m}_t)\log\left(1-\boldsymbol{m}_t\right)\right] \quad (12)$$

To avoid introducing noise from regions where the model's predictions are unreliable, we perform entropy minimization only on unlabeled pixels where the model already exhibits high confidence. The loss $\mathcal{L}_{un}$ is defined as the average entropy over unlabeled pixels whose predicted entropy is below a threshold of 0.5:

$$\mathcal{L}_{un} = \frac{\sum_{t\in\mathcal{U}} \mathcal{H}\left(\boldsymbol{m}_t\right)\cdot \mathbb{I}\left(\mathcal{H}\left(\boldsymbol{m}_t\right) < 0.5\right)}{\sum_{t\in\mathcal{U}} \mathbb{I}\left(\mathcal{H}\left(\boldsymbol{m}_t\right) < 0.5\right) + \epsilon} \quad (13)$$

where $\mathcal{U}$ denotes the set of unlabeled pixels. Meanwhile, we also compute the average entropy over labeled pixels as a weak regularization penalty, encouraging the model to make confident predictions in known regions. This loss, denoted as $\mathcal{L}_{la}$, is defined as follows:

$$\mathcal{L}_{la} = w_{weak}\cdot \frac{\sum_{t\in\mathcal{G}} \mathcal{H}\left(\boldsymbol{m}_t\right)}{|\mathcal{G}| + \epsilon} \quad (14)$$

where $\mathcal{G}$ denotes the set of labeled pixels and $|\mathcal{G}|$ is the total number of labeled pixels. The final mixed entropy loss $\mathcal{L}_{ent}$ is defined as the sum of $\mathcal{L}_{un}$ and $\mathcal{L}_{la}$. To ensure stable training, we adopt a weight ramp-up strategy, introducing a dynamic weight $\lambda(T)$ to modulate $\mathcal{L}_{ent}$.

$$\lambda(T) = w_{\max}\cdot \exp\left(-\left(1-\frac{\min\left(T, ramp\right)}{ramp}\right)^2\right) \quad (15)$$

where T denotes the current epoch and $ramp$ represents the ramp-up period. During this period, $\lambda(T)$ increases smoothly from 0 to its maximum value w_{max}. In this work, we set w_{max} and w_{weak} to 0.1, which is chosen to balance training stability and final model performance. Thus, the final loss term is defined as $\mathcal{L}_{CEM} = \lambda(T)\cdot\mathcal{L}_{ent}$.

Experiments and Results

Datasets and Implementation Details

Our experiments primarily utilize 8 mainstream benchmark datasets: NIST (Guan et al. 2019), CASIA (Dong, Wang, and Tan 2013), Columbia (Hsu and Chang 2006), Coverage (Wen et al. 2016), IMD (Novozamsky, Mahdian, and Saic 2020), CocoGlide (Nichol et al. 2021), ITW (Huh et al. 2018), and Korus (Korus and Huang 2016). We adopt the same training set split as (Zhou et al. 2024a). During training, all images are resized to 512×512. The model is trained with a batch size of 32 using the AdamW optimizer. The initial learning rate is set to 1e-4 and decayed by a factor of 0.1 every 50 epochs. Training is performed on 4 NVIDIA 4090 GPUs with a total of 70 epochs.

Comparison with SOTA Methods

Pixel-level **F1** and **AUC** are standard metrics for IML, but recent research (Ma et al. 2025) shows that **AUC** exhibits overconfidence in IML. Therefore, we evaluate all experiments using the **F1-score** with a fixed threshold of 0.5.

Image manipulation localization. Table 1 presents a comparison of all published weakly supervised IML methods to date, including WSCL (Zhai et al. 2023), EdgeCAM (Zhou et al. 2024b), SOWCL (Zhu, Li, and Wen 2025), and WSCCL (Bai 2025). While these methods rely on image-level labels, the results clearly demonstrate that our approach significantly outperforms them across all datasets, highlighting that the spatial constraints provided by scribble annotations lead to more accurate localization. Table 2 presents a comparison with several SOTA fully supervised methods, such as PCSS-Net (Liu et al. 2022), Trufor (Guillaro et al. 2023), MFI-Net (Ren et al. 2024), SparseViT (Su et al. 2025), PIM (Kong et al. 2025b), and Mesorch (Zhu et al. 2025), evaluated under both in-distribution and out-of-distribution settings. Our method not only achieves higher average performance than fully supervised baselines under standard conditions, but also demonstrates stronger generalization to previously unseen manipulation scenarios. This enhanced generalization is attributed to the balance between annotation efficiency and effective supervision achieved by scribble annotations, which provide direct spatial guidance with minimal labeling effort and help prevent overfitting to dense pixel-level labels. It is worth noting that all fully supervised methods were retrained on the same benchmarks as ours for a fair comparison. However, since among the weakly supervised methods only WSCL has publicly available code, we retrained only WSCL, while the results for the other methods are taken directly from their published papers.

Visual comparison. Fig.4 presents the segmentation results of our proposed model and several fully supervised methods under challenging scenarios, including large-scale manipulations, small-scale manipulations, and multi-object manipulations. Clearly, our method achieves the best visual performance. In large-scale manipulations, our model accurately identifies the manipulated regions, producing clear and complete segmentation, while the compared methods often suffer from false positives. In small-scale manipulations, our method provides precise localization, whereas the other methods almost entirely fail. For multi-object manipulations, our method not only detects all manipulated regions but also excels in capturing fine details.

Method	Pub.	In-Distribution (ID)					Out-of-Distribution (OOD)				
		NIST	CASIAv1	Columbia	Coverage	Avg.	CocoGlide	ITW	Korus	IMD	Avg.
PCSS-Net	TCSVT'22	0.509	0.503	0.905	0.424	0.585	0.344	0.407	0.229	0.350	0.333
Trufor	CVPR'23	0.584	0.603	0.953	0.435	0.644	0.287	0.310	0.200	0.375	0.293
IML-ViT	Arxiv'24	0.440	0.529	0.906	0.168	0.511	0.200	0.270	0.181	0.209	0.215
MFI-Net	TSCVT'24	**0.817**	0.524	0.938	0.497	0.644	0.283	0.300	0.186	0.329	0.275
SparseViT	AAAI'25	0.616	0.557	0.958	0.546	0.669	0.311	0.333	0.193	0.381	0.305
PIM	TPAMI'25	0.587	0.548	0.954	0.504	0.648	0.481	0.392	0.203	0.395	0.368
Mesorch	AAAI'25	0.802	0.703	**0.981**	0.531	0.754	0.218	0.299	0.182	0.346	0.261
Ours	-	0.721	**0.716**	0.979	**0.827**	**0.811**	**0.546**	**0.467**	**0.282**	**0.580**	**0.469**

Table 2: Comparison with other fully supervised IML methods. Bold and underlined indicate the best and second-best.

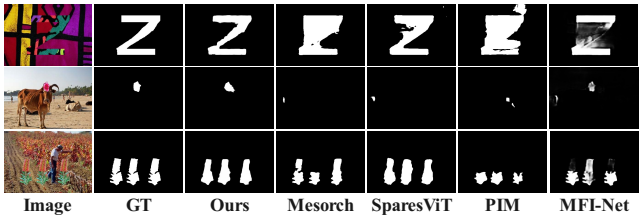

Figure 4: Visualization results of different methods.

Baseline	GAFM	PFMM	**Avg.ID**	**Avg.OOD**
✓			0.533	0.283
✓	✓		0.697	0.434
✓	✓	✓	**0.811**	**0.469**

Table 3: The ablation study for our modules.

$\mathcal{L}_{PCE}$	$\mathcal{L}_{CA}$	$\mathcal{L}_{SC}$	$\mathcal{L}_{CEM}$	**Avg.ID**	**Avg.OOD**
✓				0.627	0.394
✓	✓			0.731	0.438
✓	✓	✓		0.755	0.414
✓	✓		✓	0.733	0.418
✓	✓	✓	✓	**0.811**	**0.469**

Table 4: The ablation study for our loss functions.

Ablation Study

As shown in Table 3, introducing GAFM significantly improves performance, highlighting the importance of adaptive information regulation in feature fusion. When all modules are integrated, our model achieves the best results on both in-distribution (ID) and out-of-distribution (OOD) datasets, underscoring the synergistic effect of the proposed structural designs and feature modulation strategies. Table 4 further analyzes the contribution of each loss function. Introducing $\mathcal{L}_{CA}$ improves performance, confirming the benefit of context-aware modeling. While $\mathcal{L}_{SC}$ boosts ID results, it reduces flexibility on OOD datasets and increases prediction uncertainty, leading to performance degradation. $\mathcal{L}_{CEM}$ alone yields only marginal gains. When $\mathcal{L}_{SC}$ and $\mathcal{L}_{CEM}$ are combined, the model achieves substantial improvements on both ID and OOD datasets, indicating that structural consistency and confidence-aware entropy minimization are complementary: $\mathcal{L}_{SC}$ provides a stable structural prior, whereas $\mathcal{L}_{CEM}$ adaptively suppresses unreliable predictions in weakly annotated or unlabeled regions, thereby enhancing robustness and generalization.

Method	None	Facebook	WeiBo	WeChat	WhatsApp
MFI-Net	0.524	0.449	0.455	0.363	0.474
SparesViT	0.557	0.493	0.529	0.365	0.506
PIM	0.548	0.581	0.566	0.505	0.585
Mesorch	0.703	0.671	0.655	0.583	0.677
Ours	**0.716**	**0.685**	**0.690**	**0.623**	**0.686**

Table 5: Robustness experiments on online social networks.

Robustness and Efficiency of the Model

With the rapid development of the internet, online social platforms have become a primary channel for image dissemination. To evaluate the robustness of our model under such conditions, we followed the same benchmark as (Dong et al. 2022) and applied compression through platforms such as Facebook, Weibo, WeChat, and WhatsApp. As shown in Table 5, our model consistently maintains significant performance advantages after online transmission.

Conclusion

In this work, we present and release the first scribble-annotated IML dataset, Sc-IML, filling an important gap in weakly supervised annotation resources for the field. We also propose the first scribble-based weakly supervised IML framework, which incorporates structural consistency, prior-aware feature modulation, and gated adaptive fusion modules, significantly boosting model's robustness and localization accuracy. Moreover, the confidence-aware entropy minimization loss further enhances the model's generalization in weakly supervised and unlabeled regions. Experimental results demonstrate that our approach consistently outperforms existing fully supervised methods on both in-distribution and out-of-distribution datasets. Our work provides a new perspective for low-cost annotation and weakly supervised learning in IML, and significantly advances the development of this field.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62302427, Grant 62462060, and Grant 62472368, in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2023D01C175.

References

Bai, R. 2025. Weakly-supervised cross-contrastive learning network for image manipulation detection and localization. *Knowledge-Based Systems*, 310: 113033.

Chen, Y.; Cheng, H.; Wang, H.; Liu, X.; Chen, F.; Li, F.; Zhang, X.; and Wang, M. 2024. EAN: Edge-Aware Network for Image Manipulation Localization. *IEEE Transactions on Circuits and Systems for Video Technology*.

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*, 1–14.

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.

Gu, F.; Dai, Y.; Fei, J.; and Chen, X. 2024. Deepfake detection and localisation based on illumination inconsistency. *Int. J. Auton. Adapt. Commun. Syst.*, 17(4): 352–368.

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.

He, G.; Zhang, X.; Wang, F.; and Fu, Z. 2024. A novel copy-move detection and location technique based on tamper detection and similarity feature fusion. *Int. J. Auton. Adapt. Commun. Syst.*, 17(6): 514–529.

He, R.; Dong, Q.; Lin, J.; and Lau, R. W. 2023. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 781–789.

Hsu, J.; and Chang, S. 2006. Columbia uncompressed image splicing detection evaluation dataset. *Columbia DVMM Research Lab*, 6.

Huh, M.; Liu, A.; Owens, A.; and Efros, A. A. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 101–117.

Kong, C.; Luo, A.; Wang, S.; Li, H.; Rocha, A.; and Kot, A. C. 2025a. Pixel-Inconsistency Modeling for Image Manipulation Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

Kong, C.; Luo, A.; Wang, S.; Li, H.; Rocha, A.; and Kot, A. C. 2025b. Pixel-inconsistency modeling for image manipulation localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Korus, P.; and Huang, J. 2016. Evaluation of random field models in multi-modal unsupervised tampering localization. In *2016 IEEE international workshop on information forensics and security (WIFS)*, 1–6. IEEE.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7505–7517.

Ma, X.; Zhu, X.; Su, L.; Du, B.; Jiang, Z.; Tong, B.; Lei, Z.; Yang, X.; Pun, C.-M.; Lv, J.; et al. 2025. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. *Advances in Neural Information Processing Systems*, 37: 134591–134613.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Novozamsky, A.; Mahdian, B.; and Saic, S. 2020. IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 71–80.

Obukhov, A.; Georgoulis, S.; Dai, D.; and Van Gool, L. 2019. Gated CRF loss for weakly supervised semantic image segmentation. *arXiv preprint arXiv:1906.04651*.

Ren, R.; Hao, Q.; Niu, S.; Xiong, K.; Zhang, J.; and Wang, M. 2024. MFI-Net: Multi-Feature Fusion Identification Networks for Artificial Intelligence Manipulation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2): 1266–1280.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14318–14328.

Su, L.; Ma, X.; Zhu, X.; Niu, C.; Lei, Z.; and Zhou, J.-Z. 2025. Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7024–7032.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3): 415–424.

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.

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.

Zhou, Y.; Wang, H.; Zeng, Q.; Zhang, R.; and Meng, S. 2024a. A contribution-aware noise feature representation model for image manipulation localization. *Knowledge-Based Systems*, 111988.

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

Zhu, J.; Guo, Y.; Sun, G.; Yang, L.; Deng, M.; and Chen, J. 2023. Unsupervised Domain Adaptation Semantic Segmentation of High-Resolution Remote Sensing Imagery With Invariant Domain-Level Prototype Memory. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–18.

Zhu, X.; Ma, X.; Su, L.; Jiang, Z.; Du, B.; Wang, X.; Lei, Z.; Feng, W.; Pun, C.-M.; and Zhou, J.-Z. 2025. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11022–11030.

Zhu, Z.; Li, J.; and Wen, Y. 2025. Self-Optimization Training for Weakly Supervised Image Manipulation Localization. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.