SUMI-IFL: An Information-Theoretic Framework for Image Forgery Localization with Sufficiency and Minimality Constraints

Ziqi Sheng[1], Wei Lu[1*], Xiangyang Luo[2*], Jiantao Zhou[3], Xiaochun Cao[4]

[1]School of Computer Science and Engineering, MoE Key Laboratory of Information Technology, Guangdong Province Key Laboratory of Information Security Technology, Sun Yat-sen University.
[2]State Key Laboratory of Mathematical Engineering and Advanced Computing.
[3]State Key Laboratory of Internet of Things for Smart City, Department of Computer and Information Science, University of Macau
[4]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University.
shengzq@mail2.sysu.edu.cn, luwei3@mail.sysu.edu.cn, luoxy_ieu@sina.com, jtzhou@um.edu.mo, caoxiaochun@mail.sysu.edu.cn

Abstract

Image forgery localization (IFL) is a crucial technique for preventing tampered image misuse and protecting social safety. However, due to the rapid development of image tampering technologies, extracting more comprehensive and accurate forgery clues remains an urgent challenge. To address these challenges, we introduce a novel information-theoretic IFL framework named SUMI-IFL that imposes sufficiency-view and minimality-view constraints on forgery feature representation. First, grounded in the theoretical analysis of mutual information, the sufficiency-view constraint is enforced on the feature extraction network to ensure that the latent forgery feature contains comprehensive forgery clues. Considering that forgery clues obtained from a single aspect alone may be incomplete, we construct the latent forgery feature by integrating several individual forgery features from multiple perspectives. Second, based on the information bottleneck, the minimality-view constraint is imposed on the feature reasoning network to achieve an accurate and concise forgery feature representation that counters the interference of task-unrelated features. Extensive experiments show the superior performance of SUMI-IFL to existing state-of-the-art methods, not only on in-dataset comparisons but also on cross-dataset comparisons.

Introduction

Driven by the extensive accessibility of large-scale digital image datasets and the advancement of AIGC technologies, generating vast quantities of forgery images that surpass human detection has become remarkably effortless. However, the malicious utilization of these forged images can lead to severe consequences, such as identity theft, privacy violations, large-scale economic fraud, and the proliferation of misinformation. Given the severe consequences of image forgeries, there has been an increasing focus on developing advanced image forgery localization (IFL) technologies.

IFL is a technique that performs true/false judgment on suspicious images and further predicts tampered regions at a pixel level. Research on image forgery localization (IFL) has been proliferating and can be broadly categorized into two main groups: those focusing on extracting more comprehensive forgery clues, and those focusing on obtaining more accurate forgery clues. One line of work explores comprehensive forgery clues by designing multi-stream structures or utilizing multiple dimensions of auxiliary information (Sun et al. 2023; Liu et al. 2024). For instance, (Zhang, Li, and Chang 2024) designed a two-stream architecture incorporating RGB and frequency features to detect tampered images. MVSS-Net (Dong et al. 2022) proposed an edge-sensitive branch and noise-sensitive branch to mine the forgery edge with the aid of edge information and noise information. After fully acquiring the forgery features, some task-unrelated noises will inevitably be introduced. For example, the post-processing operation traces, the JPEG artifacts (Kwon et al. 2022a) contained in JPEG images, and reconstruction artifacts in stereo images (Luo et al. 2022). This task-irrelevant information can significantly affect the localization performance of the IFL task. Therefore, another line of work is dedicated to obtaining more accurate forgery features (Zhang et al. 2024), so as to resist the interference of task-unrelated information. For example, (Zhuo et al. 2022) utilized a self-attention mechanism including the spatial attention branch and channel attention branch to better localize forgery regions. (Li et al. 2023a) devised a region message passing controller to weaken the message passing between the forged and authentic regions, thus obtaining a refined forgery feature. Although these methods have largely advanced the IFL field, the urgent challenge of extracting forgery clues comprehensively and accurately still exists due to the rapid advancement of image forgery techniques.

To meet this challenge, we propose an information-theoretic IFL framework named SUMI-IFL employs sufficiency-view and minimality-view constraints to obtain more comprehensive and accurate forgery features. The sufficiency-view constraint is applied to the feature extraction network to ensure that the latent forgery feature contains comprehensive forgery clues by maximizing the mutual information between the latent feature and the ground-truth label. Besides, to further explore comprehensive forgery clues, we construct the latent forgery feature from several individ-

*Corresponding Author

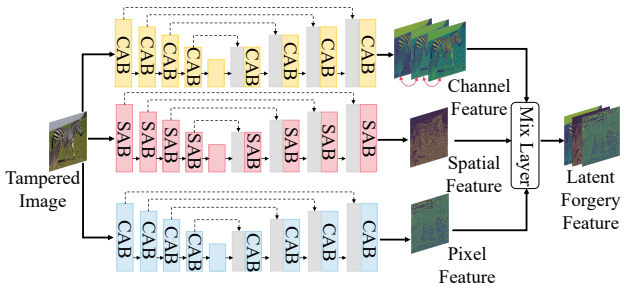

Figure 1: Illustrate the structure of the feature extraction network. We utilize three backbones to extract the channel forgery feature, the spatial forgery feature, and the pixel forgery feature of the tampered image respectively. Then the three individual image features are fed into the $\mathbf{B}_\phi$ layer to obtain the latent forgery feature. The three backbones are the U-Net structure by substituting the Conv layer with the specially designed attention blocks: channel attention block (CAB), spatial attention block (SAB), and pixel attention block (PAB).

ual forgery features from multiple perspectives. As shown in Fig. 1, we introduce three attention U-Nets to adequately extract forgery clues from different individual aspects: the channel aspect, the spatial aspect, and the pixel aspect. As for the channel aspect, forgery traces can be identified by capturing color and texture changes. For the spatial aspect, structural inconsistencies of the tampered image can be better accessed, while the pixel aspect focuses on modifications in image details. These individual forgery features focus on different aspects of the tampered image and make the fused latent forgery features more comprehensive and sufficient.

Although the latent forgery feature is able to capture sufficient forgery traces, some task-unrelated information will inevitably be introduced. Information Bottleneck (IB) theory provides a theoretical foundation for understanding the optimal trade-off between compaction and accuracy in information processing (Tishby, Pereira, and Bialek 2000). Based on the IB theory, we derive the minimality-view constraint to ensure that the final feature is concise, minimizing task-unrelated information while retaining task-related information. In particular, we map the discrete ground-truth mask to a continuous forgery feature space to guide the forgery features in eliminating task-unrelated information. With the benefit of the sufficiency-view constraint and minimality-view constraint, SUMI-IFL obtains competitive performance compared to other state-of-the-arts on both in-dataset and cross-dataset experiments. In summary, our contributions are as follows:

- We propose an innovative information-theoretic IFL framework, named SUMI-IFL, which applies sufficiency-view and minimality-view constraints to forgery feature representation, ensures the framework learns comprehensive forgery clues and counters the interference of task-unrelated features, supported by rigorous theoretical analysis.
- A sufficiency-view constraint is applied to the feature extraction network to guarantee the latent forgery feature contains comprehensive forgery clues, which is constructed by several individual forgery features.
- A minimality-view constraint is applied to the feature reasoning network to obtain the concise forgery feature by reducing task-unrelated information, thus helping the model resist the interference of task-unrelated features.

Related Work

Image Manipulation Localization

Nowadays, many researchers are making an effort to design sophisticated and complex models to achieve sound performance in image forgery localization tasks (Sheng, Yin, and Lu 2025; Wang et al. 2024). Span (Hu et al. 2020) designed a pyramid structure of local self-attention blocks to model spatial correlation in suspicious images. ObjectFormer (Wang et al. 2022) utilized an object encoder and a patch encoder to mine both the RGB features and frequency features to identify the tampering artifacts. MMFusion (Triaridis and Mezaris 2024) combined RGB images with an auxiliary forensic modality to perform image manipulation localization. MVSS-Net (Dong et al. 2022) proposed an edge-supervised branch to learn the forgery edge and a noise-sensitive branch to capture abnormal noise. In the meantime, some methods work on reasoning and fine-tuning the learned tampering features to enhance the performance. After obtaining a latent feature from a baseline detector, IF-OSN (Wu et al. 2022) further modeled the noise involved by the online social network for robust image forgery detection. CAT-Net (Kwon et al. 2022b) learned multi-scale forgery features from both the RGB stream and the DCT stream, and then all these learned features are subsequently fed into the fusion stage for a final prediction. HiFi-IFDL (Guo et al. 2023) devised a hierarchical fine-grained network to learn feature maps of different resolutions for a comprehensive representation of image forgery detection. All these works contribute a lot to the image forgery detection field. Nevertheless, challenges remain as these learned forgery features are still somewhat incomplete and redundant due to the lack of concrete theoretical guarantees. In this paper, we propose the innovative framework SUMI-IFL to explore the representation of forgery features guided by rigorous theoretical proofs.

Information Bottleneck

The concept of information bottlenecks (Tishby, Pereira, and Bialek 2000) is currently used in deep learning both theoretically and practically and provides a solid foundation to constrain the feature representation in a variety of research domains. (Li et al. 2023b) found the minimal sufficient statistics of the whole slide image and fine-tuned the backbone into a task-specific representation. IBD (Kuang et al. 2024) devised two distillation strategies that align with the two optimization processes of the information bottleneck to improve the robustness of deep neural networks. SCMVC (Cui et al. 2024) solved the issue of feature redundancy across multiple views for multi-view clustering from

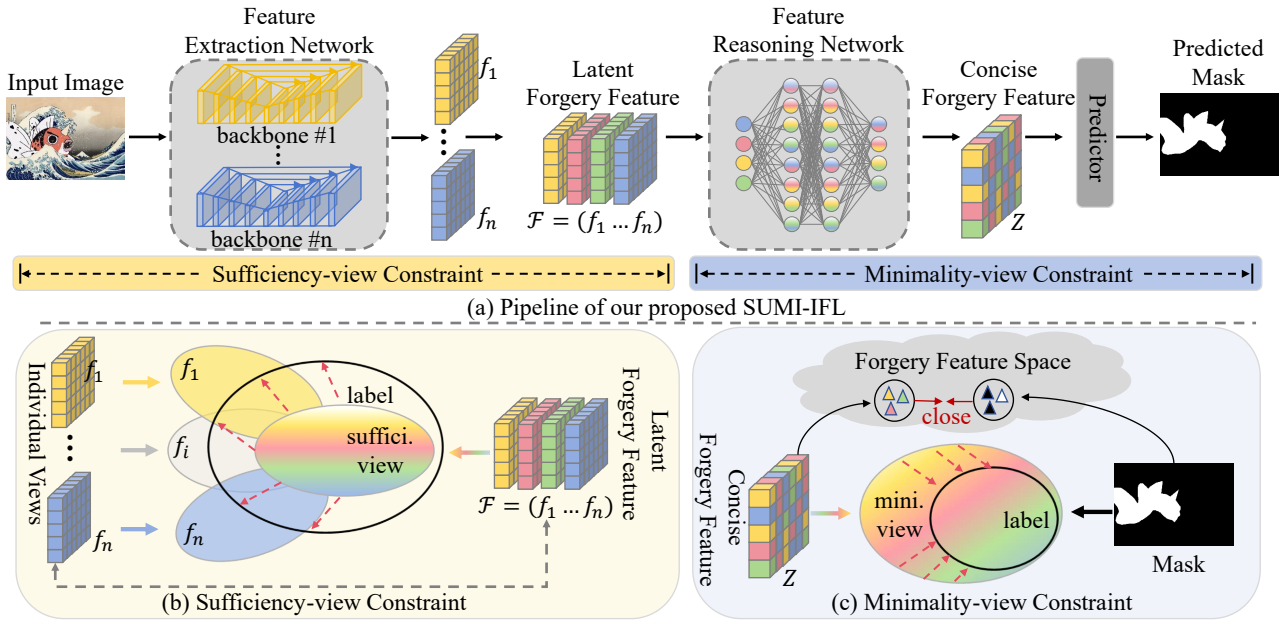

Figure 2: Overall structure of the proposed SUMI-IFL. The top part is the pipeline, which takes a suspicious image $(H \times W \times 3)$ as input, and the output is the predicted mask $(H \times W \times 1)$. The bottom parts are details of each constraint. The sufficiency-view constraint is applied to the feature extraction network to obtain a latent forgery feature, while the minimality-view constraint is applied to the feature reasoning network to get a concise forgery feature.

an information-theoretic standpoint. (Ba et al. 2024) proposed a Deepfake detection scheme to extract task-relevant local features and learn a global feature by eliminating superfluous information. Inspired by these brilliant methods, we introduce the Information Bottlenecks theory into the field of forgery image localization, constrain the representation of forgery features, and thus effectively improve localization performance.

Method

Overview

As shown in Fig. 2, we denote $h_\theta = (r \circ e)$ as a deep neural network with parameter θ, in a standard IFL task. Here, $e : \mathbb{R}^{dx} \to \mathbb{R}^{df}$ maps the inputs image X to the latent forgery feature $\mathcal{F}$, and $r : \mathbb{R}^{df} \to \mathbb{R}^{dz}$ further maps the latent feature $\mathcal{F}$ to the final predict feature Z, so that e is a feature extraction network, $\mathcal{F} = e(X)$ and r is a feature reasoning network, $r(\mathcal{F}) = r(e(X)) = Z$. Furthermore, a set of individual features from different backbones is denoted as $\mathcal{F} = e(X) = \{f_1, f_2, \dots, f_n\}$, n represent the number of backbones in the feature extraction network. Each feature has its own feature map size and channel dimension, denoted as $f_i \in \mathbb{R}^{C \times H \times W}$, where C, H and W represent the channel numbers, feature height, and width, respectively and $i = 1 \dots n$.

The sufficient-view constraint is applied to the feature extraction network e to ensure the comprehensiveness of feature representation. Specifically, we ensure the comprehensiveness of the latent forgery feature $\mathcal{F}$ by maximizing the mutual information between $\mathcal{F}$ and the ground-truth label. Besides, we uncover the independent forgery feature f_i from different perspectives to ensure that any forgery trace hidden in the tampered image is not missed.

Meanwhile, the minimality-view constraint is applied to the feature reasoning network r to guarantee that the concise forgery feature Z discards task-unrelated information while retaining task-related information. We obtain a formal representation of this constraint by deriving it from the theory of information bottleneck.

Sufficiency-View Constraint

The sufficiency-view constraint $\mathcal{L}_{SU}$ is constructed by maximizing the mutual information between $\mathcal{F}$ and the ground-truth label. In this section, we provide the key derivation of $\mathcal{L}_{SU}$ and the detailed structure of the feature extraction network.

Theoretical Proof Given a corrupted image X, we have carefully designed several feature extraction networks to extract individual forgery features $\{f_i\}_{i=1}^n$ from different perspectives. Subsequently, we employ a learnable feature fusion layer $\mathbf{B}_\phi$ to blend and reason over multiple-view forgery features to obtain the latent forgery feature $\mathcal{F}$, i.e. $\mathcal{F} = \mathbf{B}_\phi(\{f_i\}_{i=1}^n)$. Our sufficiency-view constraint objective attempts to ensure the important properties within the set $\mathcal{F} = \mathbf{B}_\phi(\{f_i\}_{i=1}^n)$, i.e., comprehensiveness. Comprehensiveness mandates the inclusion of the maximal task-related information within $\mathcal{F}$.

For the comprehensive objective, we specify the relationship between the localization label M and the latent forgery

feature $\mathcal{F}$ as follows:

$$I(M;\mathcal{F}) = I(M;f_1,\ldots,f_n) \tag{1}$$

where $I(*)$ is the mutual information. $I(M;\mathcal{F})$ represents the amount of predictive information (i.e. current task-related information) contained in $\mathcal{F}$. The comprehensiveness objective of information in $\mathcal{F}$ is given by:

$$\max[I(M;\mathcal{F})]. \tag{2}$$

Then we apply the mutual information chain rule to derive eq. (2) :

$$\begin{aligned} &\max[I(M;\mathcal{F})] = \max[I(M;f_1,\ldots,f_n)] \\ &= \max[\sum_{i=1}^{n} I(f_i;M|f_1,\ldots,f_{i-1})] \\ &\leq \max[\sum_{i=1}^{n} I(f_i;M|\mathcal{F}\setminus f_i) \end{aligned} \tag{3}$$

where $\mathcal{F}\setminus f_i = \mathbf{B}_\phi(f_1,\ldots,f_{i-1},f_{i+1},\ldots,f_n)$, $\mathbf{B}_\phi$ is a learnable feature fusion layer. For mutual information, expanding the known conditions causes the mutual information to increase or remain constant, so the inequality in Eq. (3) stands.

Nevertheless, directly estimating Eq. (3) is generally infeasible. (Poole et al. 2019) have highlighted significant challenges in mutual information estimation, chiefly attributed to the curse of dimensionality, where the number of samples required for an accurate estimate grows exponentially with the embedding dimension. To address this issue, we employ variational inference to optimize Eq. (3), bypassing the need for explicit mutual information estimation. We have the following derivation (detailed proof is in supplementary files):

$$\sum_{i=1}^{n} I(M;f_i|\mathcal{F}\setminus f_i) \geq \sum_{i=1}^{n} D_{KL}[\mathcal{P}_{\mathcal{F}}||\mathcal{P}_{\mathcal{F}\setminus f_i}], \tag{4}$$

where $\mathcal{P}_{\mathcal{F}} = p(y|\mathcal{F})$, $\mathcal{P}_{\mathcal{F}\setminus f_i} = p(y|\mathcal{F}\setminus f_i)$ represent the predicted distributions. D_{KL} denotes the Kullback-Leibler(KL) divergence.

Given the above analytical derivations, we can thus denote the sufficiency-view constraint as:

$$\mathcal{L}_{SU} = \min[\exp\left(-D_{KL}[\mathcal{P}_{\mathcal{F}}||\mathcal{P}_{\mathcal{F}\setminus f_i}]\right)] \tag{5}$$

Here, since the KL-divergence is not bounded above, i.e. $D_{KL} \in [0,\infty)$, we take the exponential of its negative value to transform the objective from maximization to minimization. The transformed objective is bounded within $(0,1]$ which is numerically advantageous. Next, the structure of the feature extraction network is elaborated to illustrate how forgery traces can be adequately extracted from different perspectives.

Feature Extraction Network The structure of the feature extraction network comprises three backbones which are based on the U-Net (Ronneberger, Fischer, and Brox 2015). All of the backbones have 5 layers of U-Net architecture with 3 blocks at each scale. The three attention backbones are constructed by substituting the Conv layer in U-Net with three novel attention blocks respectively. These attention blocks are shown in Fig. 3. The channel attention block (CAB) (Fig. 3 (a)) uses global average pooling to squeeze the input feature from C dimension to 1 dimension, then generates a channel attention map to guide the model focusing on the luminance information of tampered images. Then the spatial attention block (SAB) (Fig. 3 (b)) operates the input feature by both the global average pooling and max average pooling to squeeze dimension to 2. Because of the pooling, the output features have non-local (global) information. In other words, the SAB mainly responds to changes in global information, i.e., structure, and color information. As for the pixel attention block (PAB) (Fig. 3 (c)), it directly generates an attention map without any pooling or sampling operations, which means PAB is focused on local information of tampered images. Overall, the feature extraction network extracts forgery features, f_1, f_2, and f_3 from different individual aspects. The sufficiency-view constraint $\mathcal{L}_{SU}$ captures forgery clues from multiple and diverse perspectives, thereby reducing the risk of missing or misjudging cases and improving localization performance.

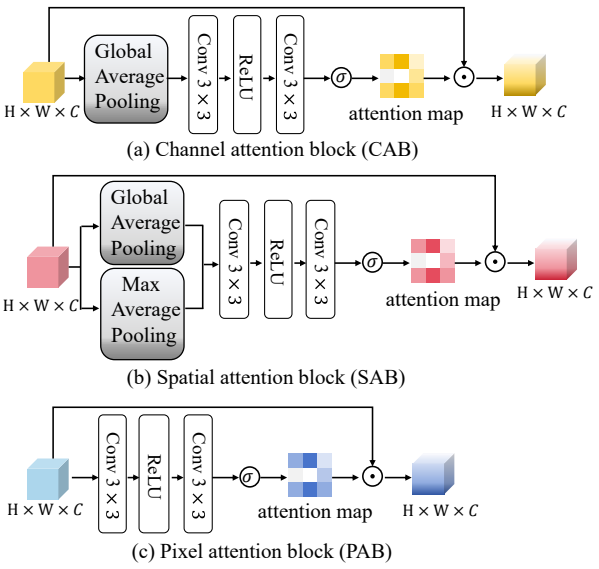

Figure 3: Illustrate the attention blocks.

Instead of adopting concatenation or addition operations, we propose a learnable feature fusion layer $\mathbf{B}_\phi$ to integrate these individual features. $\mathbf{B}_\phi$ utilizes a learnable parameter γ_ϕ to optimize the feature fusion operation through back-propagation. As a result, the fused feature $\mathcal{F}$ from $\mathbf{B}_\phi$ can be represented as:

$$\mathcal{F} = \mathbf{B}_\phi(f_1,f_2,f_3) = \frac{(1-\gamma_\phi)}{2}f_1 + \gamma_\phi f_2 + \frac{(1-\gamma_\phi)}{2}f_3. \tag{6}$$

There are also many state-of-the-art methods (Ba et al. 2024; Cui et al. 2024) dedicated to applying information theory to make feature representations as task-relevant as pos-

sible. For example, (Ba et al. 2024) attempts to reveal more forgery clues for deepfake detection tasks by extracting several orthogonal features. However, SUMI-IFL focuses on the image forgery localization task. Therefore, we do not need to make the extracted multi-view features orthogonal. Instead, these multi-view forgery features can complement each other, resulting in a more comprehensive extraction of forgery traces. Subsequently, we will apply the information bottleneck theory to eliminate task-unrelated information, which will be elaborated upon in the next section.

Minimality-View Constraint

The minimality-view constraint $\mathcal{L}_{MI}$ is derived from the theory of information bottleneck to ensure that the concise forgery feature effectively discards task-unrelated information while retaining task-related information. In this section, we provide the key derivation of minimality-view constraint $\mathcal{L}_{MI}$ and the detailed structure of the reasoning network.

Theoretical Proof After the feature extraction network, the latent forgery feature $\mathcal{F}$ already contains sufficient forgery clues but also inevitably contains task-unrelated information. Thus, we pass the latent forgery feature $\mathcal{F}$ through the reasoning network to eliminate superfluous information and obtain a concise forgery representation Z with the guidance of the minimality-view constraint. The concept of information bottlenecks (Tishby, Pereira, and Bialek 2000) is attributed to distilling superfluous noises while retaining only useful information. The information bottleneck (IB) objective can be formulated as follows:

$$\max[I(Z;M) - \beta I(\mathcal{F};Z)], \tag{7}$$

where I denotes mutual information and β controls the trade-off between the two terms. However, IB may not fully leverage the available label information, which can be crucial for improving inference performance. A study in (Fischer 2020) proposes a conditional entropy bottleneck (CEB), which enhances IB by introducing label priors in variational inference. CEB can be formulated as follows:

$$\max[I(Z;M) - \beta I(\mathcal{F};Z|M)]. \tag{8}$$

A major challenge in making the CEB practical is to estimate the mutual information accurately. We adopt the practice of variation information bottle (Alemi et al. 2016), utilizing variational inference to construct the lower bound to estimate the mutual information. Then Eq. (8) can be rewritten as :

$$\begin{aligned} &I(Z;M) - \beta I(\mathcal{F};Z|M) \\ &\geq \mathbb{E}_{p(f,m)p(z|m)}\left[\log q(m|z) - \beta \log \frac{p(z|f)}{q(z|m)}\right], \end{aligned} \tag{9}$$

where $p(z|f)$ is feature distribution, $q(m|z)$ and $q(z|m)$ are a variational approximation to the true distribution $p(m|z)$, $p(z|m)$, respectively. The detailed proof is in supplementary files. Then, the first term in Eq. (9) can be derived as:

$$\begin{aligned} &\mathbb{E}_{p(f,m)p(z|f)}\left[\log q(m|z)\right] \\ &= \mathbb{E}_{p(f)q(z|f)}\left[\int p(m|z)\log q(m|z)\, dm\right] \\ &= \mathbb{E}_{p(f)}\left[-\mathcal{L}_{CE}(q(z|f), m)\right]. \end{aligned} \tag{10}$$

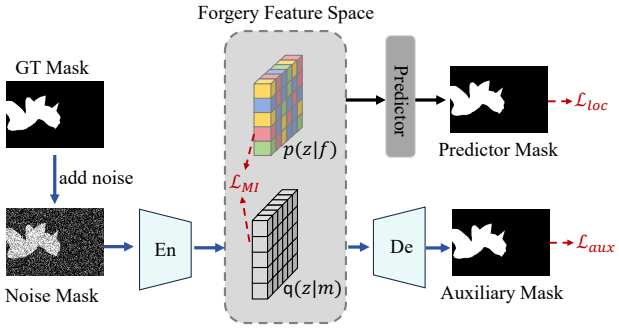

Figure 4: Illustrate the structure of the feature reasoning network.

Therefore, we obtain the localization loss $\mathcal{L}_{loc}$,

$$\mathcal{L}_{loc} = \mathcal{L}_{CE}(q(z|f), m), \tag{11}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss.

The second term in Eq. (9) can be derived as:

$$\begin{aligned} &\mathbb{E}_{p(f,m)p(z|f)}\left[\log \frac{p(z|f)}{q(z|m)}\right] \\ &= \mathbb{E}_{p(f)p(z|f)}\left[\mathrm{KL}(p(z|f)\|q(z|m))\right], \end{aligned} \tag{12}$$

Finally, we arrive at the minimality-view constraint $\mathcal{L}_{MI}$:

$$\mathcal{L}_{MI} = \mathbb{E}_{p(f)p(z|f)}\left[\mathrm{KL}(p(z|f)\|q(z|m))\right]. \tag{13}$$

Feature Reasoning Network In order to model the distribution $q(z|m)$ in the Eq. (13) of the model, we propose a mask-guided encoder-decoder structure in the feature reasoning network. As shown in Fig. 4, we first add noise to the ground truth mask. This step is because predicting the auxiliary mask from the discrete GT mask may be too simple for the encoder-decoder structure and not beneficial for training. Hinder by the box denoising training of DN-DETR (Li et al. 2022), we add the point noises to the GT mask to obtain robust models. We randomly select the points within the mask and invert the original value to represent the distinct region. In addition, we use a hyper-parameter γ to denote the noise percentage of area, so the number of noise points is $\gamma \times HW$.

Given the noise mask, we further project it to the forgery feature space to obtain the distribution $q(z|m)$ through a convolution encoder network. Therefore, the KL distance between the $p(z|f)$ and $q(z|m)$ in Eq. (13) can be easily measured by mapping the GT mask to the forgery feature space. Since $p(z|f)$ can get the predicted mask by the predictor, $q(z|m)$ can also obtain the auxiliary mask $\hat{m}$ by fed into the predictor. We can derive the auxiliary mask loss:

$$\mathcal{L}_{aux} = \mathcal{L}_{CE}(\hat{m}, m), \tag{14}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss.

Overall Objective

The total loss function $\mathcal{L}$ include four parts: the localization loss $\mathcal{L}_{loc}$, the sufficiency-view constraint $\mathcal{L}_{SU}$, the minimality-view constraint $\mathcal{L}_{MI}$, and the auxiliary mask loss L_{aux}:

$$\mathcal{L} = \mathcal{L}_{loc} + \lambda_1 \times \mathcal{L}_{SU} + \lambda_2 \times \mathcal{L}_{MI} + \lambda_3 \times L_{aux}, \tag{15}$$

where $\lambda_1 = 0.1$, $\lambda_2 = 1$, and $\lambda_3 = 0.1$.

Datasets	Tampered	Real	Training	Testing
DEFACTO-12	12000	1000	✔	✔
SSRGFD	2068	922	✔	✔
CASIAv2	5105	7491	✔	✔
Spliced COCO	917648	917648	✔	-
CIMD	100	100	-	✔
NIST16	288	288	-	✔

Table 1: Datasets used in our experiments.

Experiments

Setup

Dataset Table 1 presents the training and test datasets used in our method. We first pre-train our model on the training portions of four public datasets: DEFACTO-12 (Mahfoudi et al. 2019)(real/tampered), SSRGFD (Yin et al. 2023)(real/tampered), CASIAv2 (Dong, Wang, and Tan 2013) (real/-tampered), and Spliced COCO (Kwon et al. 2022b) created by CAT-Net (Kwon et al. 2022b) based on the COCO 2017 dataset (Lin et al. 2014). Then we test our model on the testing portions of the above datasets, except Spliced COCO.

To further evaluate the generalization capability of SUMI-IFL, we also compare the localization performance on two other datasets: CIMD (Zhang, Li, and Chang 2024) (real/-tampered) and NIST16 (Guan et al. 2019) (real/tampered). All forgery images are cropped into 256×256 patches. To evaluate the localization performance of the proposed SUMI-IFL, following the previous method (Rao et al. 2022), we adopt the F1 score and Area Under Curve (AUC) as the evaluation metric.

Implementation Details The proposed SUMI-IFL is implemented with PyTorch and all experiments are performed on the NVIDIA GTX GeForce A100 GPU platform. The whole model is trained with batch size 12 for 100 epochs with AdamW optimizer, and an initial learning rate of 5e-4 set by cosine annealing scheduler, weight decay as 0.005.

Comparison With State-of-the-Art Methods

We compare SUMI-IFL with other state-of-the-art methods under three settings: 1) in-dataset comparisons: training on the compound forgery dataset and evaluating on the comprehensive test datasets. 2) cross-dataset comparisons: directly applying the pre-trained model on an unseen dataset to assess generalization. 3) robustness evaluation: applying JPEG compression and Gaussian blur to the test dataset to evaluate robustness. We evaluate the performance with the seven state-of-the-art methods: MMFusion (Triaridis and Mezaris 2024), EITL-Net (Guo, Zhu, and Cao 2024), HiFi-IFDL (Guo et al. 2023) , WSCL (Zhai et al. 2023), IF-OSN (Wu et al. 2022), MVSS-Net (Dong et al. 2022), PSCC-Net (Liu et al. 2022).

In-Dataset Comparisons Table 2 reports the optimal and suboptimal localization in terms of F1 score and AUC score. We can observe that SUMI-IFL achieves the highest performance on DEFACTO, SSRGFD, and CASIAv2 datasets. In particular, SUMI-IFL achieves a 0.7995 F1 score on the stereo forgery dataset SSRGFD and outperforms the suboptimal method by 15.7%. This confirms that the minimality-view constraint can help the framework capture accurate forgery traces even with the interference from reconstruction artifacts present in the SSRGFD dataset. In the other two datasets, SUMI-IFL also outperforms the other methods in terms of both F1 scores and AUC scores, demonstrating its capability to obtain a superior representation of forgery features.

Methods	DEFACTO-12		SSRGFD		CASIAv2	
	F1	AUC	F1	AUC	F1	AUC
MMFusion	0.8052	0.9056	0.5305	0.7864	0.5837	0.6928
EITL-Net	0.8189	0.9381	0.5842	0.8269	0.5281	0.7581
HiFi-Net	0.2235	0.4654	0.0977	0.5112	0.3496	0.5605
WSCL	<u>0.8395</u>	0.9487	0.6471	0.8611	<u>0.7347</u>	<u>0.8941</u>
IF-OSN	0.8258	<u>0.9504</u>	<u>0.6734</u>	<u>0.8866</u>	0.6867	0.8583
MVSS-Net	0.7709	0.9373	0.5439	0.8538	0.5175	0.7781
PSCC-Net	0.4846	0.6188	0.3817	0.4626	0.4381	0.5221
SUMI-IFL	**0.9249**	**0.9760**	**0.7995**	**0.9493**	**0.7604**	**0.8984**

Table 2: In-dataset comparisons of manipulation localization in terms of F1 score and AUC scores. The first and second rankings are shown in **bold** and <u>underlined</u> respectively.

Methods	CIMD		NIST16	
	F1	AUC	F1	AUC
MMFusion	0.1293	0.5486	<u>0.5519</u>	<u>0.6470</u>
EITL-Net	0.0215	**0.5598**	0.3246	0.6901
HiFi-Net	0.1049	0.5222	0.4021	0.5681
WSCL	0.0734	0.6273	0.3136	0.6177
IF-OSN	0.0426	0.5369	0.4326	0.6417
MVSS-Net	0.0114	0.476	0.3181	0.7017
PSCC-Net	<u>0.1707</u>	0.4404	0.5039	0.5153
SUMI-IFL	**0.1738**	<u>0.5513</u>	**0.6178**	**0.7339**

Table 3: Cross-dataset comparisons of manipulation localization in terms of F1 score and AUC scores. The first and second rankings are shown in **bold** and <u>underlined</u> respectively.

Cross-Dataset Comparisons To further demonstrate the generalizability of SUMI-IFL, we utilize two test datasets with completely different distributions from the training datasets. Table 3 reports the cross-dataset performance in terms of F1 score and AUC score, SUMI-IFL consistently ranks among the top two in the test datasets. The CIMD is a newly published dataset with relatively small tampered regions, which is a challenge for IFL methods. In this dataset, the localization performance decreases for all IFL methods. However, SUMI-IFL can learn comprehensive forgery clues and outperforms other IFL methods. In the NIST16 dataset, the proportion of tampered images is higher. Although all methods demonstrate strong performance, the F1 score of SUMI-IFL exceeds the second-best method by 9.7%. The performance of cross-dataset comparisons demonstrates the sound generalization of SUMI-IFL.

Robustness Evaluation We apply different image distortion methods on raw images from the DEFACTO-12 dataset and evaluate the robustness of our SUMI-IFL. The distortion types include 1) JPEG compression with a fixed quality factor and 2) Gaussian blurring with a fixed kernel size. We compare the manipulation localization performance (AUC

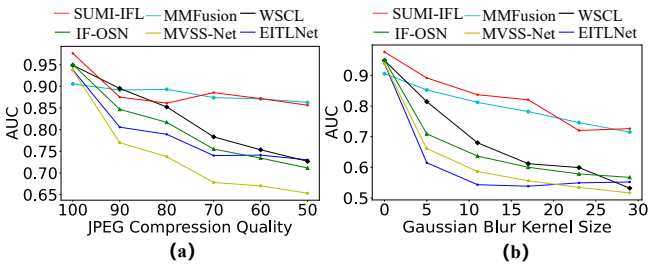

Figure 5: Robust evaluation against JPEG compression and Gaussian Blurs on DEFACTO.

ID	LOSS			DEFACTO-12		SSRGFD	
	$\mathcal{L}_{SU}$	$\mathcal{L}_{MI}$	$\mathcal{L}_{aux}$	F1	AUC	F1	AUC
1	-	✔	✔	0.8335	0.8888	0.6783	0.7841
2	✔	-	✔	0.8774	0.9523	0.6784	0.7942
3	✔	✔	-	0.9010	0.8972	0.7215	0.8765
4	✔	✔	✔	**0.9249**	**0.9760**	**0.7995**	**0.9493**

Table 4: Abalation study of the proposed $\mathcal{L}_{SU}$ and $\mathcal{L}_{MI}$ in terms of F1 score and AUC scores. The bold mark best performance

scores) of our pre-trained models with other methods on the distorted dataset and present the results in Fig. 5. As shown in Fig. 5 (a), under JPEG compression the performance degradation of SUMI-IFL is lower than the other baselines, indicating that the proposed method has good JPEG robustness. As illustrated in Fig. 5 (b), SUMI-IFL can also resist Gaussian blur, indicating that the proposed method is robust against low-quality images.

Ablation Study

In this section, we study the effect of removing the sufficiency-view constraint $\mathcal{L}_{SU}$, the minimality-view constraint $\mathcal{L}_{MI}$, and the auxiliary mask loss $\mathcal{L}_{aux}$. We train models on the compound datasets mentioned before and test them on the test portions of the DEFACTO-12 and SSRGFD datasets. The absence of either loss leads to a significant drop in model performance. Quantitatively, $\mathcal{L}_{SU}$ and $\mathcal{L}_{MI}$ have a dominant contribution to our method, resulting in an F1 increase of 9.8% and 5.1% on DEFACTO-12 and SSRGFD, respectively. Without $\mathcal{L}_{aux}$, the F1 score drops 2.5% and 9.8% on DEFACTO and SSRGFD, respectively. This empirical evidence suggests that the incorporation of the proposed losses results in extracting more comprehensive and less task-unrelated forgery features, facilitating the subsequent localization performance.

Visualization Results

As shown in Fig. 6, we provide predicted forgery masks of various methods. It can be observed that some methods incorrectly identify certain image objects as tampered regions, such as in the third row of the first column, where MMFusion mistakenly identifies the lower-right area of the image as tampered. The comparison of visualization results demonstrates that SUMI-IFL can not only locate the tampered regions more accurately but also produce clearer regions. This is attributed to the sufficiency-view constraint $\mathcal{L}_{SU}$ and the minimality-view constraint $\mathcal{L}_{MI}$, which enable the model to obtain comprehensive task-related feature representations while effectively resisting the interference of task-unrelated features.

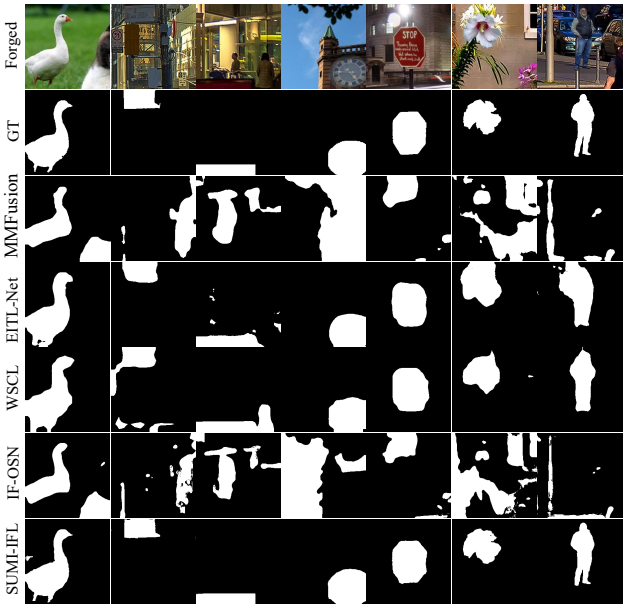

Figure 6: Visualization of the predicted manipulation mask by different methods. From top to bottom, we show forged images, GT masks, predictions of MMFusion, EITL-Net, WSCL, IF-OSN, and SUMI-IFL.

Conclusion

In this paper, we proposed a novel information-theoretic IFL framework, SUMI-IFL, that leverages sufficiency-view constraints and minimality-view constraints to constrain the representation of forgery features. In one respect, the sufficiency-view constraint is applied to the feature extraction network, guaranteeing the latent forgery features capture comprehensive task-related information. The feature extraction network consists of three attention backbones to uncover forgery clues from different perspectives. In another aspect, the minimality-view constraint is employed in the feature reasoning network, assuring the concise forgery feature to eliminate superfluous information thus helping the model to resist the interference of the redundancy feature. We provided a detailed derivation of these two constraints based on the theories of mutual information maximization and information-theoretic bottlenecks, respectively. The superior performance of SUMI-IFL is demonstrated by extensive experimental results obtained across several benchmark tests, demonstrating that the two critical constraints contribute to a more comprehensive and accurate feature representation.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62072480).

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Ba, Z.; Liu, Q.; Liu, Z.; Wu, S.; Lin, F.; Lu, L.; and Ren, K. 2024. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 719–728.

Cui, C.; Ren, Y.; Pu, J.; Li, J.; Pu, X.; Wu, T.; Shi, Y.; and He, L. 2024. A novel approach for effective multi-view clustering with information-theoretic perspective. *Advances in Neural Information Processing Systems*, 36.

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

Dong, J.; Wang, W.; and Tan, T. 2013. Casia image tampering detection evaluation database. In *IEEE China Summit and International Conference on Signal and Information Processing*, 422–426.

Fischer, I. 2020. The conditional entropy bottleneck. *Entropy*, 22(9): 999.

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 *IEEE Winter Applications of Computer Vision Workshops*, 63–72.

Guo, K.; Zhu, H.; and Cao, G. 2024. Effective image tampering localization via enhanced transformer and co-attention fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4895–4899.

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

Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; and Nevatia, R. 2020. SPAN: Spatial pyramid attention network for image manipulation localization. In *The 16th European Conference on Computer Vision*, 312–328.

Kuang, H.; Liu, H.; Wu, Y.; Satoh, S.; and Ji, R. 2024. Improving adversarial robustness via information bottleneck distillation. *Advances in Neural Information Processing Systems*, 36.

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

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

Li, D.; Zhu, J.; Wang, M.; Liu, J.; Fu, X.; and Zha, Z.-J. 2023a. Edge-aware regional message passing controller for image forgery localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8222–8232.

Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L. M.; and Zhang, L. 2022. Dn-DETR: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13619–13627.

Li, H.; Zhu, C.; Zhang, Y.; Sun, Y.; Shui, Z.; Kuang, W.; Zheng, S.; and Yang, L. 2023b. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7454–7463.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *The 13th European Conference on Computer Vision*, 740–755.

Liu, H.; Tan, Z.; Tan, C.; Wei, Y.; Wang, J.; and Zhao, Y. 2024. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10770–10780.

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.

Luo, J.; Liu, L.; Xu, W.; Yin, Q.; Lin, C.; Liu, H.; and Lu, W. 2022. Stereo super-resolution images detection based on multi-scale feature extraction and hierarchical feature fusion. *Gene Expression Patterns*, 45: 119266.

Mahfoudi, G.; Tajini, B.; Retraint, F.; Morain-Nicolier, F.; Dugelay, J. L.; and Pic, M. 2019. DEFACTO: Image and Face Manipulation Dataset. In *27th European Signal Processing Conference*, 1–5.

Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; and Tucker, G. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, 5171–5180.

Rao, Y.; Ni, J.; Zhang, W.; and Huang, J. 2022. Towards JPEG-Resistant Image Forgery Detection and Localization Via Self-Supervised Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–12.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention*, 234–241.

Sheng, Z.; Yin, C.; and Lu, W. 2025. Exploring multi-scale forgery clues for stereo super-resolution image forgery localization. *Pattern Recognition*, 161: 111230.

Sun, Z.; Jiang, H.; Wang, D.; Li, X.; and Cao, J. 2023. SAFL-Net: Semantic-agnostic feature learning network with auxiliary plugins for image manipulation detection. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, 22424–22433.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

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

Wang, H.; Wang, J.; Hu, X.; Hu, B.; Yin, Q.; Luo, X.; Ma, B.; and Sun, J. 2024. Detecting Double Mixed Compressed Images Based on Quaternion Convolutional Neural Network. *Chinese Journal of Electronics*, 33(3): 657–671.

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

Wu, H.; Zhou, J.; Tian, J.; Liu, J.; and Qiao, Y. 2022. Robust image forgery detection against transmission over online social networks. *IEEE Transactions on Information Forensics and Security*, 17: 443–456.

Yin, C.; Zhang, B.; Luo, J.; Zhu, C.; and Lu, W. 2023. SSRGFD: stereo super-resolution general forensic dataset. *Journal of Image and Graphics*, 28(11): 3386–3399.

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

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

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

Zhuo, L.; Tan, S.; Li, B.; and Huang, J. 2022. Self-adversarial training incorporating forgery attention for image forgery localization. *IEEE Transactions on Information Forensics and Security*, 17: 819–834.