UnionFormer: Unified-Learning Transformer with Multi-View Representation for Image Manipulation Detection and Localization

Shuaibo Li[1,2] Wei Ma[1†] Jianwei Guo[2] Shibiao Xu[3] Benchong Li[1] Xiaopeng Zhang[2]
[1]Beijing University of Technology
[2]MAIS, Institute of Automation, Chinese Academy of Sciences
[3]Beijing University of Posts and Telecommunications

Abstract

We present UnionFormer, a novel framework that integrates tampering clues across three views by unified learning for image manipulation detection and localization. Specifically, we construct a BSFI-Net to extract tampering features from RGB and noise views, achieving enhanced responsiveness to boundary artifacts while modulating spatial consistency at different scales. Additionally, to explore the inconsistency between objects as a new view of clues, we combine object consistency modeling with tampering detection and localization into a three-task unified learning process, allowing them to promote and improve mutually. Therefore, we acquire a unified manipulation discriminative representation under multi-scale supervision that consolidates information from three views. This integration facilitates highly effective concurrent detection and localization of tampering. We perform extensive experiments on diverse datasets, and the results show that the proposed approach outperforms state-of-the-art methods in tampering detection and localization.

1. Introduction

The rapid progression of deep generative models, such as GANs [21, 43, 60], VAEs [31, 50], and Diffusion Models [10, 45, 53], has facilitated the widespread availability of Artificial Intelligence Generated Content (AIGC) tools [8]. At the same time, image editing tools have become exceptionally user-friendly and powerful, capable of creating highly realistic images and videos. This assists users in better expressing their creativity but also intensifies the malicious use of editing techniques to tamper with multimedia content, resulting in the proliferation of faked images on the Internet [57]. Therefore, developing a universally effective method to discern the authenticity of images and accurately locate the modified regions has become crucial. Research on related algorithms has become a hot topic [3, 28], and many state-of-the-art methods based on deep learning models have been proposed.

Digital image tampering falls into three main categories [19]: splicing, which involves copying regions from one image to another; copy-move, entailing the copying or moving of elements within the same image; and removal, the process of erasing parts of an image and creating visual consistent content to obscure the alterations. These manipulations leave traces between the tampered regions and their surroundings, causing inconsistencies between the authentic and forgery regions. Unlike traditional detection or segmentation tasks emphasizing high-level semantic information, image tampering detection prioritizes local semantic-agnostic clues that distinguish authenticity rather than semantic content. Therefore, the critical challenge in tampering detection is learning generalizable features that combine different level information and capture multiple scale inconsistencies between authentic and tampered areas. Previous methods primarily utilized deep convolutional neural networks designed for high-level visual tasks as feature encoders or directly connected features from different layers [23, 27, 40, 71], which could not adequately represent tampering traces. Inspired by [9, 12, 67], we designed a Boundary Sensitive Feature Interaction Network (BSFI-Net) specifically for extracting forensics artifacts and integrated it as the feature encoder in our framework. BSFI-Net is a parallel CNN-Transformer structure that can reinforce edge responses while effectively interacting between local features and global representations to explore consistencies within images at different scales.

On the other hand, many tampering artifacts imperceptible in the RGB view become distinctly noticeable in the noise view. Employing fixed [18] or learnable high-pass filters [6, 35, 66] to convert RGB images into noise maps can suppress content and highlight the low-level forgery clues. Thus, developing a multi-view strategy that simultaneously models the RGB and noise dimensions is essential to detect subtle tampering traces. Our framework adopts a dual-stream architecture to independently construct representation for RGB and noise views, subsequently merging

†Corresponding author.

them to enhance discriminative capability and generalizability. Furthermore, we incorporate contrastive supervision to improve the collaboration between the two views.

In addition, to create spatially coherent and semantically consistent images, tamper operations invariably alter entire objects to conceal evidence, namely performing object-level manipulation. Current advanced methods focus on pixel or patch-level consistencies, overlooking object-level information. Conversely, we argue that image manipulation detection should extend beyond merely identifying out-of-distribution pixels or patches to also capture the anomalies in object consistency and distribution resulting from manipulation. Due to hyper-realistic tampered images generated by diffusion models [4, 5, 20, 30, 44, 65, 69], leveraging object view information becomes particularly crucial. Diffusion-based models [4, 30, 44] repeatedly update initial noise across the image, enhancing spatial continuity and leaving fewer RGB and noise traces. Moreover, unlike authentic image sources, auto-generated forgery portions guided by natural language prompts are more likely to exhibit object incongruities. Recent Diffusion models [20, 29, 55, 64] have attempted to solve this issue by employing object-centric approaches, underscoring the necessity and feasibility of object view clues for tampering detection. However, creating and integrating such a novel view with others for tampering artifact representation presents a significant challenge, requiring new architectures and learning strategies.

Considering the above vital points, we introduce UnionFormer, a unified-learning transformer framework with multi-view representation for image manipulation detection and localization, as illustrated in Figure 1. Firstly, we use BSFI-Net as the feature encoder to obtain the generalizable features under RGB and noise views and combine them. Then, we utilize the fused features to conduct a unified learning process, which includes three sub-tasks: object consistency modeling, forgery detection, and localization. In unified learning, our model establishes the object view representation and integrates three view information into a unified manipulation discriminative representation (UMDR) to simultaneously accomplish forgery detection and localization. To summarize, our main contributions are as follows:

- We propose UnionFormer, a novel image forensics transformer framework. By employing unified learning with multi-scale supervision, the UnionFormer integrates information from all three views to execute image manipulation detection and localization simultaneously.
- We introduce BSFI-Net, a hybrid network structure for superior artifact representation learning, which enhances boundary response while revealing local inconsistencies at different levels across domains.
- With the unified learning of UMDR, we construct an innovative object view representation capable of capturing the inconsistency among objects and aggregated information from three views for forgery detection.
- We involve comprehensive experiments across various benchmarks, demonstrating that our method attains state-of-the-art results in both detection and localization tasks.

2. Related Work

Forgery Artifacts Representation. Most early works [17, 33, 42] design hand-crafted features to characterize tampering traces, often detecting specific types of manipulation. However, in real-world scenarios, various editing operations are usually combined, and the types are unknown, promoting more work to focus on practical general tampering detection [13, 23, 27, 59, 62]. Achieving general detection requires more generalizable and semantic-agnostic features, so a series of works explore clues beyond the RGB view to capture a broader range of tampering traces. The most common approach is to use fixed [18] or learnable [6, 34, 66] filters to transform the image into the noise view to highlight weak low-artifacts. Some other works leverage frequency-aware clues to provide a complementary viewpoint [49, 54]. These low-level features are always combined with the high-level features from the RGB view for more effective detection [23, 27, 34, 36, 62, 70]. For instance, [13] employs dual attention to combine information from RGB and noise views. [59] extracts high-frequency features of the images and combines them with RGB features as multimodal patch embedding. In contrast, we not only combine tampering representations from both streams (RGB and noise views) but also facilitate their sufficient interaction through contrastive supervision. Moreover, we incorporate a novel view that models the inconsistencies between objects, providing robust additional cues for manipulation detection.

Transformer in Vision. Transformer [58] employs self-attention mechanisms to model long-range dependencies, and it has been widely successful in natural language processing (NLP). Some works are inspired to explore the use of transformer architecture for various computer vision tasks and showed superior performance. Specifically, ViT [16] reshapes images into patch sequences and feeds them into a transformer encoder for image classification. DETR [9] and Deformable DETR [72] implement end-to-end object detection using a transformer encoder-decoder architecture with learnable queries and bipartite matching. CMX [68] proposed a transformer framework for semantic segmentation that integrates RGB and other modal information. In this work, we first introduce a CNN-Transformer parallel encoder, BSFI-Net, for tampering feature extraction. Then, we utilize a unified-learning transformer framework to integrate multiple views information for image manipulation

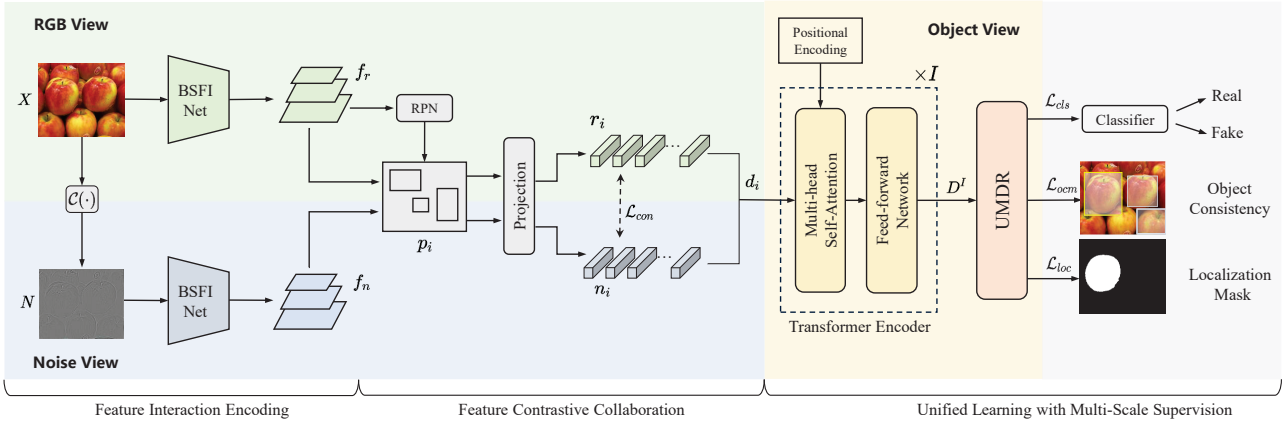

Figure 1. An overview of UnionFormer. We achieve simultaneous tampering detection and localization by integrating tampering clues from three view representations, with each view represented by a different color background. We obtain representations under the RGB and noise views through BSFI-Net and construct the object view representation based on both in the unified learning. Meanwhile, information from all three views is interactively fused into a unified manipulation discriminative representation (UMDR) for detection and localization.

detection and localization.

3. Method

In this section, we first provide an overview of UnionFormer and a detailed introduction to each component. We aim to fully leverage rich artifacts from three views for simultaneous tampering detection and localization. We achieve this through a unified learning process under multi-scale supervision. As illustrated in Figure 1, input RGB image X is firstly transformed into a noise view representation $N = \mathcal{C}(X)$ using constrained CNN [7], which can reveal low-level tampering. Then, both X and N are individually fed into the Boundary Sensitive Feature Interaction Networks (BSFI-Net) for feature encoding. High-frequency edge features (H) are incorporated with either X or N as inputs into the BSFI-Net to boost edge responsiveness. This allows us to acquire generalizable and discriminative features under the RGB and noise views, constructing two feature pyramids $f_r = \mathcal{E}_1(X, H)$, $f_n = \mathcal{E}_2(N, H)$. Subsequently, we use a Region Proposal Network (RPN) [51] to obtain a set of Regions of Interest (RoIs), represented as p_i, from the feature f_r. RoI information is extracted from f_r and f_n, then flattened to get embedding representations for proposals, denoted as r_i, n_i. The RGB feature r_i and noise feature n_i for each proposal are concatenated to generate the fused proposal feature di, which is input into the I transformer Encoder layer.

During the unified learning phase, we address three subtasks: modeling object consistencies, binary classification of authenticity, and tampered region localization. After the transformer encoder, the forgery-discriminative query embeddings D^I are fed into the unified manipulation discriminative representation part to generate three predictions for three sub-tasks. As shown in Figure 1, we employ multi-scale supervision with a unified form for three sub-tasks, including $\mathcal{L}_{cls}$, $\mathcal{L}_{ocm}$, and $\mathcal{L}_{loc}$.

3.1. Feature Interaction Encoding

RGB and Noise View Representation. We utilize a dual-stream structure to harness clues from both RGB and noise views in the feature encoding stage. The RGB stream is designed to capture visually apparent tampering artifacts, while the noise stream aims to explore the distribution inconsistencies between tampered and genuine regions. We employ the learnable constrained convolutional layer proposed in [7] to transform the RGB image into the noise view.

As noted in Section 2, the edges of tampered regions and their surroundings exhibit more prominent tampering clues. Therefore, we enhance high-frequency edge information in both streams to concentrate the network's response on tampered regions. Specifically, we utilize the Discrete Cosine Transform (DCT) to convert the image data X into the frequency domain and then apply a high-pass filter to obtain the high-frequency component. We then convert the high-frequency component back to the spatial domain to facilitate feature interaction and preserve local consistency. Thus, we get the edge-enhanced information H as follows:

$$H = \mathcal{T}_d^{-1}\left(\mathcal{F}_h\left(\mathcal{T}_d(X), \beta\right)\right), \tag{1}$$

where $\mathcal{T}_d$ represents DCT, $\mathcal{F}_h$ represents the high-pass filter, and β is the threshold. We input X and N separately into the BSFI-Net, along with H for feature encoding, as illustrated in Figure 2.

Boundary Sensitive Feature Interaction Network. In addition to enhancing boundary responses, integrating local features and global representations is crucial for image forgery detection. This allows for a comprehensive analysis

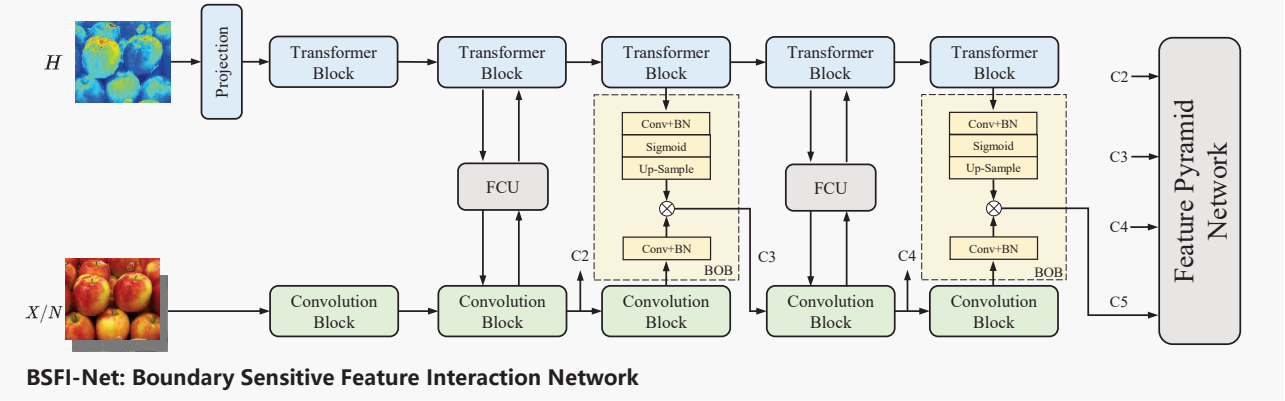

Figure 2. An overview of BSFI-Net. FCU represents the Feature Coupling Unit, and BOB represents the Boundary Oriented Block.

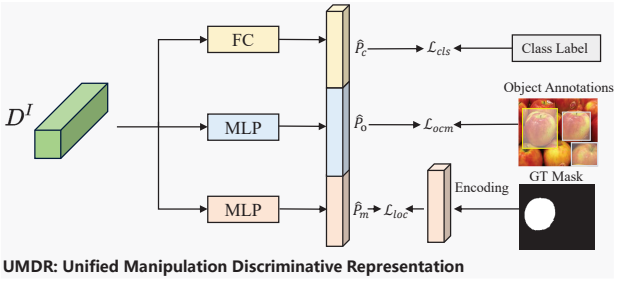

Figure 3. The learning of UMDR with multi-scale supervision.

of inconsistencies within the image at various scales. Inspired by [48], we propose a CNN-Transformer concurrent network called BSFI-Net, which maintains edge sensitivity while facilitating thorough interaction between features at different scales in the two branches.

As shown in Figure 2, the CNN branch serves as the main branch, taking an RGB or noise image as input to encode local information. The transformer branch, with input as edge enhancement information H, guides the CNN branch to focus on tampered regions and transmits long-distance inconsistencies between image patches to it. We use the Feature Coupling Unit (FCU) proposed by [48] to eliminate the misalignment between feature maps from the CNN branch and patch embeddings from the transformer branch. Moreover, we design a Boundary Oriented Block (BOB) to facilitate transmitting high-level patch consistency and boundary information from the transformer branch to the CNN branch, guiding the latter.

The CNN branch consists of five convolution blocks, similar to the ResNet construction [24]. Like [16, 48], the transformer branch consists of 5 repeated transformer blocks, consisting of a multi-head self-attention module and an MLP block. The same tokenization operation as ViT [16] is adopted. In FCU, 1×1 convolution and re-sampling are used to align channels and spatial dimensions before adding patch embeddings and CNN features. In BOB, feature maps from the CNN branch are fed into a 1×1 convolution layer, a batch normalization layer, a sigmoid layer, and up-sampled to high resolution by bilinear interpolation. Then, the features from the CNN branch are subjected to an element-wise multiplication with the long-distance discriminate weights. We pre-train BSFI-Net as a feature encoder to generate RGB and noise view representation, and two feature pyramids f_r, f_n are produced by the Feature Pyramid Network [38] based on the intermediate feature maps $\{C2, C3, C4, C5\}$. The training details are provided in Section 4.1.

3.2. Feature Contrastive Collaboration

In the feature collaboration stage, inspired by [51, 56], we first employ a Region Proposal Network (RPN) based on the RGB feature pyramid f_r to generate a set of Regions of Interest (RoIs). Then, we utilize RoIAlign [25] to extract the information of RoIs from the feature pyramids f_r and f_n of two streams. In addition to feature concatenation, we employ contrastive supervision to promote collaboration between two views. We treat the tampered proposals from different streams as positive proposals, and the tampered proposals and authentic proposals are assigned as negative pairs. Following the InfoNCE loss [47, 67], the contrast loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{con}} = &-\frac{1}{N}\sum_i \log\frac{\exp(s_0)}{\exp(s_0)+\sum_j \exp(s_1)} \\ &-\frac{1}{N}\sum_i \log\frac{\exp(s_0)}{\exp(s_0)+\sum_j \exp(s_2)},\end{aligned} \tag{2}$$

where s_0 stands for the similarity between positive pairs, s_1 denotes the similarity between RGB tampered embeddings and noise authentic embeddings, and s_2 signifies the similarity between RGB authentic embeddings and noise tampered embeddings. The contrastive loss $\mathcal{L}_{con}$ is introduced into the supervision of unified learning and will be discussed in Section 3.3.

3.3. Unified Learning with Multi-Scale Supervision

Transformer Encoder. Our Unified learning module is an encoder-only transformer architecture that processes the fused proposal embeddings d_i, along with their specific positional encoding as input. Within each layer of the transformer encoder, self-attention mechanisms aggregate the information across different proposal embeddings and capture their long-distance dependencies, implying object consistencies. In detail, we utilize a transformer decoder featuring six layers, a width of 512, and eight attention heads. The feedforward network (FFN) within the transformer has a hidden size 2048. After the transformer encoder, we generate the discriminative query embeddings D^I, fed into the unified manipulation discriminative representation (UMDR) part to generate predictions for three sub-tasks, viz. object consistency modeling, image manipulation detection, and localization.

Unified Manipulation Discriminative Representation. After the transformer encoder, each tampering discriminative query in D^I represents the tampering clues across three views of the corresponding proposal. Figure 3 shows the learning process of three sub-tasks. UMDR is learned under the supervision of authenticity classification, object consistency modeling, and manipulation localization branches. The same as DETR [9] and SOLQ [12], the classification branch is a fully connected (FC) layer to predict the authenticity confidences $\hat{P}_c$. The object consistency modeling branch is a multi-layer perception (MLP) with a hidden size of 256 to predict object spatial information $\hat{P}_o$. The manipulation localization branch is also a multi-layer perception with a hidden size of 1024 to predict localization mask vector $\hat{P}_m$. The supervision for the first two branches is similar to DETR[9]. In the third branch, we employ the mask vector, obtained by encoding the ground truth mask, as the supervision information. During the inference process, the compressed encoding procedure is applied to $\hat{P}_m$ for reconstructing the localization mask. In the compression encoding, we utilize Principal Component Analysis (PCA) to transform 2D spatial binary masks into 1D mask vectors.

Loss Function. The overall loss function for supervision of the UnionFormer can be expressed as:

$$\mathcal{L}_{union} = \lambda_{cls} \cdot \mathcal{L}_{cls} + \mathcal{L}_{ocm} + \lambda_{loc} \cdot \mathcal{L}_{loc} + \beta \cdot \mathcal{L}_{con}, \quad (3)$$

where $\mathcal{L}_{cls}$ denotes the focal loss [39] for classification. $\mathcal{L}_{loc}$ denotes the L_1 loss for localization mask vector supervision. $\mathcal{L}_{con}$ is the contrastive learning loss introduced in Section3.2. λ_{cls}, λ_{loc}, and β are the corresponding modulation coefficients. The $\mathcal{L}_{ocm}$ is the loss for object consistency modeling, which is defined as:

$$\mathcal{L}_{\text{ocm}} = \lambda_{L_1} \cdot \mathcal{L}_{L_1} + \lambda_{gious} \cdot \mathcal{L}_{gious}, \quad (4)$$

where $\mathcal{L}_{L_1}$ and $\mathcal{L}_{gious}$ are L_1 loss and generalized IoU loss [52], which is the same as DETR. λ_{L_1} and λ_{gious} are corresponding coefficients. Following [12], $\mathcal{L}_{loc}$ is not included in the bipartite matching process.

4. Experiments

4.1. Experimental Setup

Training. We used a large-scale training dataset including various types of tampered and authentic images. It is divided into five sections: 1) CASIA v2 [14], 2) Fantastic Reality [32], 3) Tampered COCO, derived from COCO 2017 datasets [37], 4) Tampered RAISE, constructed based on the RAISE dataset [11], and 5) Pristine images selected from the COCO 2017 and RAISE datasets. We randomly add Gaussian noise or apply JPEG compression to the synthetic data to simulate the visual quality and tampering traces in realistic scenarios. During the training process, we sequentially train BSFI-Net, RPN, and the entire UnionFormer in three stages.

Testing. To comprehensively evaluate and compare our model with various state-of-the-art methods, we utilized six publicly available testing datasets and one more dataset of hyper-realistic tampered images created by the Blended Diffusion model [4]. Specifically, we employed CASIA v1 [14], Columbia [26], Coverage [61], NIST16 [22], IMD20 [46] and CocoGlide [23]. Then, we construct BDNIE, including 512 hyper-realistic fake images we generated from the advanced blended Diffusion model for text-driven natural image editing. The details of the training and testing data are provided in the Supplementary.

Evaluation Metric. We evaluated the performance of the proposed method in the task of image tampering detection and localization. For the task of localizing image manipulations, we report the pixel-level Area Under Curve (AUC) and F1 score, using both the best and the fixed 0.5 thresholds. For the detection task following [23], we adopt image-level AUC and balanced accuracy, which considers both false alarms and missed detection, in which case the threshold is set to 0.5. To ensure fairness and accuracy in the comparison, some result values for other methods are taken from the literature [23, 59].

Implementation Details. The BSFI-Net is trained with cross-entropy loss for 100 epochs, employing the AdamW optimizer [41], with a batch size of 512 and a weight decay of 0.05. The initial learning rate is set to 0.001 and decays in a cosine schedule.

During the training of complete UnionFormer with $\mathcal{L}_{union}$, inspired by [56, 63], we adopt a 36-epoch ($3\times$) schedule to train the Unionformer for 2.7×10^5 iterations with batch size 16. An AdamW optimizer is also utilized in this stage. The learning rate is set to 10^{-4} at the beginning and multiplied by 0.1 at 1.8×10^5 and 2.4×10^5 iterations.

Method	Optimal threshold						Fixed threshold (0.5)					
	Columbia	*Coverage*	*CASIA v1*	*NIST16*	*CoCoGlide*	*AVG*	*Columbia*	*Coverage*	*CASIA v1*	*NIST16*	*CoCoGlide*	*AVG*
ManTra-Net [62]	0.650	0.486	0.320	0.225	0.673	0.471	0.508	0.317	0.180	0.172	0.516	0.339
SPAN [27]	0.873	0.428	0.169	0.363	0.350	0.437	0.759	0.235	0.112	0.228	0.298	0.326
MVSS-Net [13]	0.781	0.659	0.650	0.372	0.642	0.621	0.729	0.514	0.528	0.320	0.486	0.515
PSCC-Net [40]	0.760	0.615	0.670	0.210	0.685	0.588	0.604	0.473	0.520	0.113	0.515	0.445
CAT-Net v2 [34]	0.923	0.582	0.852	0.417	0.603	0.675	0.859	0.381	0.752	0.308	0.434	0.547
TruFor [23]	0.914	**0.735**	0.822	0.470	0.720	0.732	0.859	**0.600**	0.737	0.399	0.523	0.624
Ours	**0.925**	0.720	**0.863**	**0.489**	**0.742**	**0.748**	**0.861**	0.592	**0.760**	**0.413**	**0.536**	**0.632**

Table 1. Performance of pixel-level F1 with optimal and fixed threshold for image manipulation localization task.

4.2. Comparision with state-of-the-art

Baseline. To ensure a fair and accurate comparison, we only selected state-of-the-art methods for which authors provided pre-trained models, released source code, or evaluated under a common criterion [27, 40, 59]. To reduce biases, we exclusively considered the methods or versions trained on the datasets that do not overlap with the test datasets. In detail, we included seven state-of-the-art methods: Mantra-Net [62], SPAN [27], PSCC-Net [40], MVSS-Net [13], CAT-Net v2 [34], ObjectFormer [59], and TruFor [23].

Localization Results. Table 2 and Table 1 present the results of image tampering localization based on pixel-level AUC and F1 score metrics, respectively. The top-ranking method is denoted in bold, a horizontal line represents the second-ranking method, and the same annotation is applied in Table 4 and Table 3. Our method demonstrates the best performance across all datasets for pixel-level AUC evaluation. As for F1 evaluation, our method ranks the best or second best across all datasets. On average, we achieved a notable advantage, regardless of using an optimal or fixed threshold. In fact, on the relatively novel CocoGlide dataset, which includes diffusion-based local manipulations, we outperform the second-placed TruFor by 2.2% and 1.3% on the two thresholds, respectively. This is due to UnionFormer constructing object view artifacts expression, which can reveal inconsistencies between regions generated with diffusion models and authentic areas. These comparisons indicate that our method possesses strong generalization and a superior ability to capture tampering artifacts.

Detection Results. Table 4 indicates the comparative results for tampering detection. Following [23], we use the maximum value of the localization map as the detection statistic for methods not explicitly designed for the detection task. UnionFormer achieves optimal performance on all datasets except Columbia and demonstrates marked superiority in average results, whether measured by AUC or balanced accuracy. As mentioned in [13, 23], accuracy is sensitive to threshold selection and challenging to determine without a well-calibrated dataset. However, our method and the second-placed TruFor have achieved commendable results in this demanding scenario. We maintain a 2.5% and 2% lead in the average AUC and accuracy, respectively. This advantage is primarily attributed to the unified learning process of our framework. Unified learning typically facilitates the mutual enhancement of localization and detection tasks. The model's performance is further enhanced as both sub-tasks are mastered through a unified manipulation discriminative representation.

Method	Columbia	Coverage	CASIA v1	NIST16	IMD20	AVG
ManTra-Net [62]	0.824	0.819	0.817	0.795	0.748	0.801
SPAN [27]	0.936	0.922	0.797	0.840	0.750	0.849
PSCC-Net [40]	0.982	0.847	0.829	0.855	0.806	0.864
ObjectFormer [59]	0.955	0.928	0.843	0.872	0.821	0.884
TruFor [23]	0.947	0.925	0.957	0.877	-	0.927
Ours	**0.989**	**0.945**	**0.972**	**0.881**	**0.860**	**0.929**

Table 2. Performance of pixel-level AUC for image manipulation localization task. The results of Trufor on IMD20 are not reported because IMD20 is included in its training datasets.

Distortion	SPAN	PSCC-Net	ObjectFormer	Ours
w/o distortion	0.8359	0.8547	0.8718	**0.8813**
Resize(0.78×)	0.8324	0.8529	0.8717	**0.8726**
Resize(0.25×)	0.8032	0.8501	0.8633	**0.8719**
GSBr($k = 3$)	0.8310	0.8538	0.8597	**0.8651**
GSB($k = 15$)	0.7915	0.7993	0.8026	**0.8430**
GSN($\sigma = 3$)	0.7517	0.7842	0.7958	**0.8285**
GSN($\sigma = 15$)	0.6728	0.7665	0.7815	**0.8057**
JPEG($q = 100$)	0.8359	0.8540	0.8637	**0.8802**
JPEG($q = 50$)	0.8068	0.8537	0.8624	**0.8797**

Table 3. AUC scores for the localization performance on the NIST 16 dataset.

Robustness Evaluation. We tested the robustness of UnionFormer by applying image distortion to NIST 16 dataset images. Following [40, 59], we included four types of distortions: 1) changing the size of images to different scales; 2) applying Gaussian blur with a kernel size k; 3) adding Gaussian noise characterized by a standard deviation σ; 4) applying JPEG compression to the images, utilizing a quality factor q. We compare the pixel-level AUC performance with other methods. Table 3 show that our method exhibits robustness to various distortion operations, outperforming others.

Method	Image-level AUC						Accuracy					
	Columbia	*Coverage*	*CASIA v1*	*NIST16*	*CoCoGlide*	*AVG*	*Columbia*	*Coverage*	*CASIA v1*	*NIST16*	*CoCoGlide*	*AVG*
ManTra-Net [62]	0.810	0.760	0.644	0.624	0.778	0.723	0.500	0.500	0.500	0.500	0.500	0.500
SPAN [27]	**0.999**	0.670	0.480	0.632	0.475	0.651	0.951	0.605	0.487	0.597	0.491	0.626
MVSS-Net [13]	0.984	0.733	0.932	0.579	0.654	0.776	0.667	0.545	0.808	0.538	0.536	0.619
PSCC-Net [40]	0.300	0.657	0.869	0.485	0.777	0.618	0.508	0.550	0.683	0.456	0.661	0.572
CAT-Net v2 [34]	0.977	0.680	0.942	0.750	0.667	0.803	0.803	0.635	0.838	0.597	0.580	0.691
TruFor [23]	0.996	0.770	0.916	0.760	0.752	0.839	**0.984**	0.680	0.813	0.662	0.639	0.756
Ours	0.998	**0.783**	**0.951**	**0.793**	**0.797**	**0.864**	0.979	**0.694**	**0.843**	**0.680**	**0.682**	**0.776**

Table 4. Performance of image-level AUC and balanced accuracy for image manipulation detection.

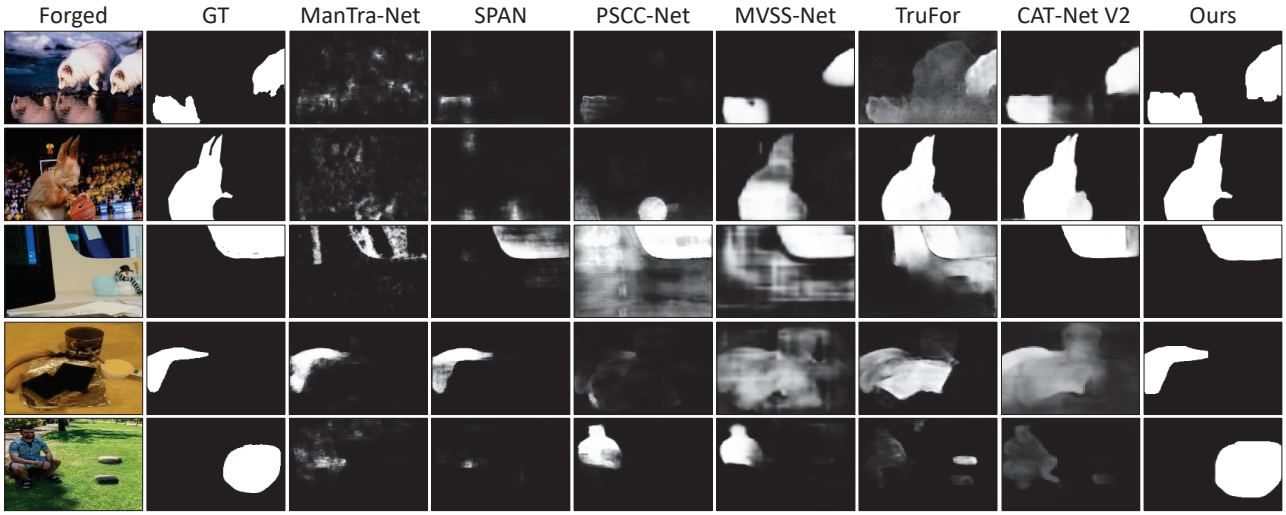

Figure 4. Qualitative comparison results. The first to fourth rows are respectively sourced from CASIA v1 [14], Columbia [26], Coverage [61], and IMD20 [46]. The last row is from the BDNIE dataset.

Variant Models	CASIA v1		NIST 16	
	AUC	F1	AUC	F1
RGB View (baseline)	0.778	0.701	0.724	0.423
RGB+Noise Views	0.865	0.767	0.807	0.448
RGB +Noise+Object Views (w/o L_{con})	0.950	0.849	0.853	0.472
UnionFormer (w/ ResNet)	0.895	0.786	0.826	0.453
UnionFormer (Ours)	**0.972**	**0.863**	**0.881**	**0.489**

Table 5. Ablation results on CASIA and NIST16 datasets.

4.3. Visualization Results

Qualitative Comparison. Figure 4 presents localization results across various datasets. Our method can accurately locate tampered regions, predicting more detailed and clear boundaries. This is due to our multi-view artifacts capture and BSFI-Net, where frequency information boosts edge response, and the interactions between branches enhance the generalization and discrimination of features. Thanks to the modeling of object view clues and the unified learning framework, our method achieves satisfactory results on the challenging BDNIE dataset, while other methods fail.

Visualization of Different View Representation. In Figure 5, we visualize noise features and the edge-guided features of the transformer branch in BSFI-Net. As shown in columns 1 to 4, some images may appear natural in the RGB view, but their tampered/authentic parts are readily distinguished in the frequency domain or under noise view. Columns five and six show the RGB features generated by a single CNN branch and the dual branch of BSFI-Net. Compared to using only the CNN branch, BSFI-Net more accurately activates the tampered regions, thanks to edge guidance and long-distance clues provided by the transformer branch.

Furthermore, we quantitatively analyze the object view, as shown in Figure 6. We derive the affinity matrix A_i from the transformer encoder during the unified learning phase. Based on A_i, we randomly select a subset of proposal embeddings and compute their average affinity with other proposals, denoted as e_i. e_i is then normalized to the range $[0, 1]$ and used as a color coefficient to visualize proposals, with

$\mathcal{L}_{\mathbf{ocm}}$	AUC	$\lambda_{\mathbf{loc}}$	AUC	$\mathbf{n_v}$	AUC	Type	AUC
w/	0.881	0.5	0.802	144	0.824	Sparse	0.847
w/o	0.796	1	0.881	256	0.881	DCT	0.860
-	-	2	0.836	400	0.813	PCA	0.881

Table 6. Ablation results for the UMDR on the NIST 16 dataset, where "n_v" denotes the dimension of the mask vector, and "Type" indicates the type of compression coding used.

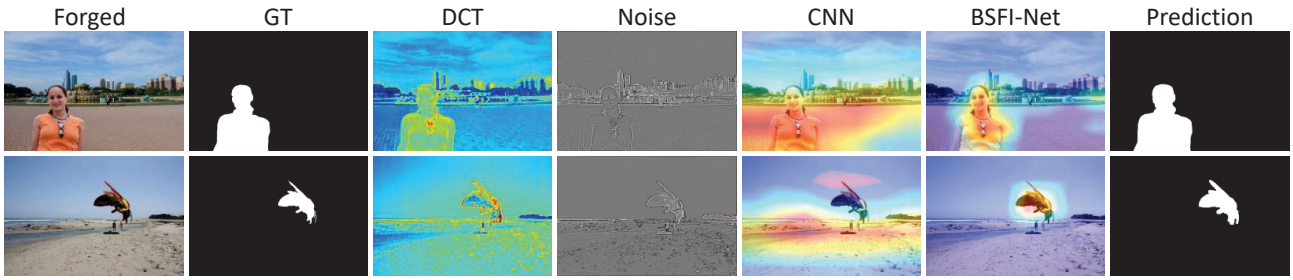

Figure 5. Visualization of diverse features. From left to right, we display the forged image, reference mask, edge-guided input of BSFI-Net, noise view input, CAM of the feature maps from CNN and BSFI-Net, and the prediction mask of UnionFormer.

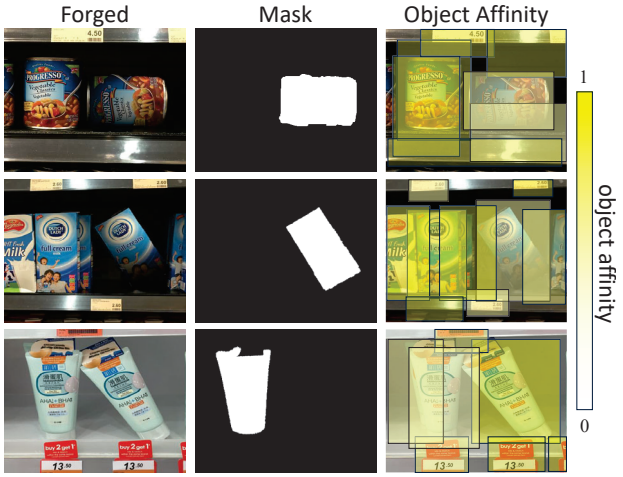

Figure 6. Visualization of object view representation. From left to right, we display the forged image, ground truth mask, and the visualization of object affinity.

lighter colors indicating lower affinity. The results show that proposals with forged objects have a lower average affinity with other regions, demonstrating UMDR's ability to capture inconsistencies between real and fake objects.

4.4. Ablation Study

Ablation studies were carried out to assess the impact of critical components within our approach. The quantitative results are listed in Table 5. We can observe that by adding noise stream on the first baseline model, the AUC scores increase by 8.7% on CASIA v1 and 8.3% on NIST 16, while further adding object view representation, the AUC scores continue to increase by 10.7% on CASIA v1 and 7.4% on NIST 16. This demonstrates the effectiveness of noise and object view representations. Moreover, when contrastive supervision is lacking, or BSFI-Net is replaced with ResNet-50 [24], the model's performance experiences a significant decline. This highlights the efficacy of the interaction between the two streams and the exceptional capability of the BSFI-Net in characterizing forgery artifacts.

The BOB and FCU modules within BSFI-Net improve the interaction between its two branches and can effectively eliminate feature misalignment between them. When BOB or FCU is removed individually, the overall model's localization AUC scores on the NIST 16 dataset decrease by 4.8% and 6.3% respectively. We further conduct experiments to investigate the effect of several key factors in UMDR, viz. λ_{loc}, $\mathcal{L}_{ocm}$, the mask vector dimension n_v, and the type of compression coding. We compare three compression coding methods: Sparse Coding [15], Discrete Cosine Transform (DCT) [2], and Principal Component Analysis (PCA) [1]. As shown in Table. 6, when equipped with contrastive loss, using PCA as the encoding type, and setting λ_{loc} and $\mathcal{L}_{ocm}$ to 1 and 256 respectively, the model performs the best on the NIST 16 dataset.

5. Conclusion

In this paper, we introduced UnionFormer, a unified-learning transformer framework that leverages clues from three distinct views for image manipulation detection and localization. UnionFormer employs BSFI-Net as a feature encoder to extract highly discriminative features under RGB and noise views. Then, through a unified learning process with three tasks, UnionFormer models the discontinuity between objects, i.e., object view representation, and learns a unified discriminative representation. The unified representation integrating information from three views has strong generalizability and discrimination. It can accurately identify various image manipulations, whether traditional manual editing or natural language-driven tampering based on diffusion models. Moreover, the unified learning framework enables the mutual enhancement of sub-tasks, achieving high-precision detection and localization. Comprehensive experiments conducted on various datasets demonstrate the efficacy of the proposed method.

Acknowledgements. We thank the anonymous reviewers for their valuable suggestions. This work is funded by the National Natural Science Foundation of China (Nos. 62176010, 61771026, U21A20515, 62172416, 62376271), and the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2022131).

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987. 8

[2] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. 8

[3] Saadaldeen Rashid Ahmed, Emrullah Sonuç, Mohammed Rashid Ahmed, and Adil Deniz Duru. Analysis survey on deepfake detection and recognition with convolutional neural networks. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–7. IEEE, 2022. 1

[4] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022. 2, 5

[5] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 2

[6] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10, 2016. 1, 2

[7] Belhassen Bayar and Matthew C Stamm. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 13(11):2691–2706, 2018. 3

[8] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023. 1

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2, 5

[10] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[11] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, page 219–224, New York, NY, USA, 2015. Association for Computing Machinery. 5

[12] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. *Advances in Neural Information Processing Systems*, 34: 21898–21909, 2021. 1, 5

[13] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022. 2, 6, 7

[14] 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, 2013. 5, 7

[15] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. 8

[16] 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. 2, 4

[17] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5):1566–1577, 2012. 2

[18] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 1, 2

[19] Oran Gafni and Lior Wolf. Wish you were here: Context-aware human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7840–7849, 2020. 1

[20] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Xingqian Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. PAIR-Diffusion: A Comprehensive Multimodal Object-Level Image Editor. *arXiv e-prints*, art. arXiv:2303.17546, 2023. 2

[21] 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

[22] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N. Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 63–72, 2019. 5

[23] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 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*, pages 20606–20615, 2023. 1, 2, 5, 6, 7

[24] 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. 4, 8

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 4

[26] Yu-feng Hsu and Shih-fu Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *2006 IEEE International Conference on Multimedia and Expo*, pages 549–552, 2006. 5, 7

[27] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 312–328. Springer, 2020. 1, 2, 6, 7

[28] Suhaib Wajahat Iqbal and Bhavna Arora. Machine learning techniques for image manipulation detection: A review and analysis. In *The International Conference on Recent Innovations in Computing*, pages 209–224. Springer, 2022. 1

[29] Jindong Jiang, Fei Deng, Gautam Singh, and Sungjin Ahn. Object-centric slot diffusion. *arXiv preprint arXiv:2303.10834*, 2023. 2

[30] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[32] Vladimir V. Kniaz, Vladimir Knyaz, and Fabio Remondino. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 5

[33] Neal Krawetz and Hacker Factor Solutions. A picture's worth. *Hacker Factor Solutions*, 6(2):2, 2007. 2

[34] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. 2, 6, 7

[35] Haodong Li and Jiwu Huang. Localization of deep inpainting using high-pass fully convolutional network. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8310, 2019. 1

[36] Shuaibo Li, Shibiao Xu, Wei Ma, and Qiu Zong. Image manipulation localization using attentional cross-domain cnn features. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5614–5628, 2023. 2

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 5

[38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 4

[39] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 5

[40] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. 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, 2022. 1, 6, 7

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[42] Babak Mahdian and Stanislav Saic. Using noise inconsistencies for blind image forensics. *Image and vision computing*, 27(10):1497–1503, 2009. 2

[43] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[44] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2

[45] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[46] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2020. 5, 7

[47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[48] 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 (ICCV)*, pages 367–376, 2021. 4

[49] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103. Springer, 2020. 2

[50] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015. 3, 4

[52] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 5

[53] 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. 1

[54] Zenan Shi, Xuanjing Shen, Hui Kang, and Yingda Lv. Image manipulation detection and localization based on the

dual-domain convolutional neural networks. *IEEE Access*, 6: 76437–76453, 2018. 2

[55] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 2

[56] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3611–3620, 2021. 4, 5

[57] Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, 6(1):2056305120903408, 2020. 1

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[59] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022. 2, 5, 6

[60] Zhengwei Wang, Qi She, and Tomas E Ward. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021. 1

[61] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage — a novel database for copy-move forgery detection. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 161–165, 2016. 5, 7

[62] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019. 2, 6, 7

[63] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[64] Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models. *arXiv preprint arXiv:2305.11281*, 2023. 2

[65] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 2

[66] Chao Yang, Huizhou Li, Fangting Lin, Bin Jiang, and Hao Zhao. Constrained r-cnn: A general image manipulation detection model. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 1, 2

[67] Yuyuan Zeng, Bowen Zhao, Shanzhao Qiu, Tao Dai, and Shu-Tao Xia. Towards effective image manipulation detection with proposal contrastive learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 4

[68] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 2

[69] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023. 2

[70] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[71] Peng Zhou, Bor-Chun Chen, Xintong Han, Mahyar Najibi, Abhinav Shrivastava, Ser-Nam Lim, and Larry Davis. Generate, segment, and refine: Towards generic manipulation segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13058–13065, 2020. 1

[72] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2