Towards Generic Image Manipulation Detection with Weakly-Supervised Self-Consistency Learning

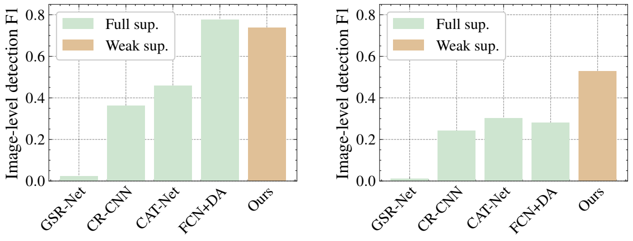
Yuanhao Zhai   Tianyu Luan   David Doermann   Junsong Yuan
University at Buffalo
{yzhai6, tianyulu, doermann, jsyuan}@buffalo.edu

Abstract

As advanced image manipulation techniques emerge, detecting the manipulation becomes increasingly important. Despite the success of recent learning-based approaches for image manipulation detection, they typically require expensive pixel-level annotations to train, while exhibiting degraded performance when testing on images that are differently manipulated compared with training images. To address these limitations, we propose weakly-supervised image manipulation detection, such that only binary image-level labels *(authentic or tampered with) are required for training purpose. Such a weakly-supervised setting can leverage more training images and has the potential to adapt quickly to new manipulation techniques. To improve the generalization ability, we propose weakly-supervised self-consistency learning (WSCL) to leverage the weakly annotated images. Specifically, two consistency properties are learned: multi-source consistency (MSC) and inter-patch consistency (IPC). MSC exploits different content-agnostic information and enables cross-source learning via an online pseudo label generation and refinement process. IPC performs global pair-wise patch-patch relationship reasoning to discover a complete region of manipulation. Extensive experiments validate that our WSCL, even though is weakly supervised, exhibits competitive performance compared with fully-supervised counterpart under both in-distribution and out-of-distribution evaluations, as well as reasonable manipulation localization ability.*

1. Introduction

With the advent of increasingly powerful image editing techniques [39, 24, 26, 56, 36, 20, 59, 58, 61, 20, 44], image manipulation has never been so convenient, and can be easily accomplished using natural language [44, 20, 44] or sketch [39, 59, 58, 61] by general users. Such advances allow malicious users to easily manipulate images, creating fake news, promoting blackmail, and generating Deepfakes [60, 48]. Thus, detecting the authenticity of an image is crucial for media forensics and credible information sharing.



(a) In-distribution (IND) manipulation detection result on the testing split of the CASIA dataset [10, 11].

(b) Out-of-distribution (OOD) manipulation detection results, which are averaged over Columbia [17], Coverage [51] and IMD2020 [33].

Figure 1. Image-level manipulation detection performance comparison with existing fully-supervised methods [65, 57, 22, 6]. All methods are trained on CASIA [10, 11]. Our weakly-supervised method achieves comparable performance with fully-supervised methods under both IND and OOD manipulation detections.

Despite previous efforts to detect image manipulations, existing solutions still confront several challenges when dealing with real problems. First, although learning-based image manipulation techniques demonstrate excellent performance compared with traditional methods, they may not easily generalize well to testing images that are manipulated differently compared with training images. As more sophisticated image manipulation techniques continue to emerge, it is exceedingly challenging, if not impracticable, to encompass all manipulation methods in the training data to enable effective handling of novel manipulations. As shown in Fig. 1, despite work well in the training image dataset, the performance of learning-based methods can degrade considerably when testing on out-of-distribution images, *i.e*., the unknown unknowns. Moreover, most learning-based methods for detecting image manipulation rely on the full supervision, *i.e*., training with pixel-level mask [42, 25, 55, 57, 65, 6, 9]. This approach is commonly adopted due to the creation of image manipulation datasets using sophisticated software [17, 11, 51, 15], where manipulated images and masks are generated simultaneously. Although the pixel-mask can provide full supervision to help the model differentiate authentic and tampered image regions, the cost of such image annotation is non-trivial and limits the amount of the training images. On the other hand, emerging language-driven image editing/synthesis or sketch-based manipulation methods do not necessarily generate pixel-level masks during the editing process [39, 59, 58, 61, 44, 20, 44, 36], but still have great potential to help train image manipulation detection if properly used.

To address the limitations of previous fully-supervised image manipulation detection methods, we propose weakly-supervised image manipulation detection (W-IMD), where only binary image-level labels are required to tell whether a given image is authentic or tampered with, thereby eliminating the need for pixel-level masks during training. We observe that image manipulation detection typically relies on inconsistency detection between features of the manipulated regions compared to features from the surrounding regions. Thus, we propose two different self-consistency learning schemes: (1) multi-source consistency (MSC) and (2) inter-patch consistency (IPC) to achieve weakly-supervised self-consistency learning (WSCL) that aims to improve the generalization ability of image manipulation detection.

For (1) multi-source consistency (MSC) learning, we take advantage of content-agnostic information by using different noise patterns in the image [14, 2] in a late-fusion manner. Specifically, we build an exclusive model on different sources (*e.g*., raw RGB image, and its noise maps) and generate an ensemble prediction by averaging predictions from different models. Intuitively, each source may focus on different locations, and locations where all models have consistent high/low activations and are likely to be manipulated/authentic. Hence, we use the ensemble prediction as a pseudo ground truth to guide each individual model, and enable them to learn cross-source information. When combining predictions from different sources, the ensemble model can be more reliable and accurate than single models. For (2) the inter-patch consistency (IPC) learning, it learns global pair-wise image patch-patch similarities in a self-supervised learning manner. By learning the pair-wise relationship, IPC helps the model to differentiate low-level authentic and tampered image patch features. Enforcing the IPC constraints helps to correct potential false positives, estimate a more complete image region of manipulation, and mitigate overfitting.

We conducted experiments on seven benchmark datasets to validate the effectiveness of our weakly-supervised method. First, we follow the conventional setting of image manipulation detection and demonstrate that our WSCL achieves comparable image-level detection performances with several fully-supervised methods under both in-distribution and out-of-distribution evaluations. Furthermore, we validate that our method can be easily extended to new manipulations where only image-level labels are available. By fine-tuning on the image-level labels, our WSCL achieves even better performance. Finally, we also demonstrate that our method achieves reasonable pixel-level manipulation localization performance even under the setting of weakly-supervised learning.

To summarize, our contributions are threefold.

- We first propose weakly-supervised image manipulation detection (W-IMD), where only binary image-level labels are required to achieve image manipulation detection. Such a paradigm eliminates the need for pixel-level annotations and can be easily adapted to new mask-free image editing techniques.
- We propose weakly-supervised self-consistency learning (WSCL) for the W-IMD task. By exploiting the multi-source consistency and inter-patch consistency, our WSCL learns and fuses information from different content-agnostic sources, performs global image patch-patch relationship learning, and promotes generic image manipulation detection. Our WSCL also has the capability to locate the manipulation region in the pixel-level[1].
- Experiments validate that our WSCL achieves strong in-distribution and out-of-distribution image manipulation detection capability, promising results when fine-tuned with image-level labels on novel manipulations, and reasonable manipulation localization ability.

2. Related Work

Image manipulation detection. We focus on detecting three types of image manipulation, *i.e*., copy-move [7, 41, 51, 53, 54], splicing [8, 19, 52, 42], and inpainting [68]. Specifically, copy-move denotes copying and pasting image content within the same image, and splicing indicates pasting image content from one image to another. Inpainting removes a particular area from an image and fills the region with new content estimated from the surrounding.

Traditional unsupervised methods detect manipulations by exploiting specific low-level tampering artifacts, such as color filter array (CFA) analysis [13], local noise analysis [29], and double JPEG compression [3]. However, these methods assume that all given images consist of authentic and tampered pixels, and perform two-class clustering on pixels to locate the manipulation. Thus, they detect manipulation out of all testing images, meaning that they always achieve 0 specificity and 1 sensitivity. Recent fully-supervised methods exploit content-agnostic features to lo-

[1] In this paper, we use "detection" to indicate image-level fake/authentic classification, and use "localization" to indicate pixel-level manipulation localization.

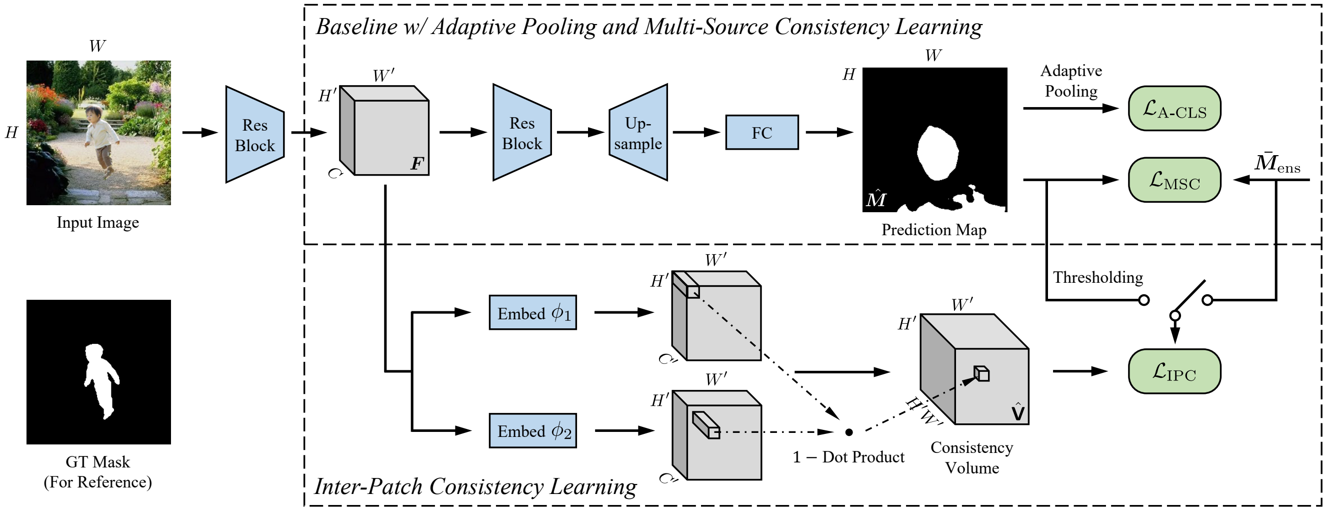


Figure 2. The single-stream overview. Given an input image, a baseline method (upper) predicts a manipulation map. The prediction map is supervised by an adaptive pooling-based classification loss $\mathcal{L}_{\text{A-CLS}}$ and a multi-source consistency learning loss $\mathcal{L}_{\text{MSC}}$. Meanwhile, the inter-patch consistency branch (bottom) learns a consistency volume measuring patch-patch similarities. The consistency volume is supervised by the inter-patch consistency loss $\mathcal{L}_{\text{IPC}}$.

calize manipulations [66, 25, 55, 18, 57, 6, 9], given the hypothesis that manipulation areas differ from pristine parts in terms of their noise distributions. Two of the noise filters are the most popular, *i.e.*, the steganalysis rich model (SRM) filter [14, 66] and the Bayar convolutional filters [2]. Specifically, SRM filters [14, 66] use predefined kernels to learn different types of noise residuals among the neighboring pixels of the center pixel, followed by linear or non-linear max/min operations. The Bayar convolutional filters [2] improve the SRM filters by using learnable weights, with the constraint that the weighted sum of neighboring pixels equals the negative of the weight of the center pixel. In addition to the SRM and Bayar filters, CAT-Net [22] learns compression artifacts from the RGB and DCT domains jointly. MVSS-Net [6] learns semantic-agnostic information by exploiting noise distribution and boundary artifacts in a multi-view, multi-scale fashion. Except for leveraging noise maps, GSR-Net [65] designs a pipeline to automatically generate copy-move images to enhance the training set. Mantra-Net [55] learns the manipulation trace by conducting fine-grained manipulation type classification.

There exist several works that exploit the image consistency [4, 19, 30, 31, 32, 63] for image forensics and Deepfake detection. And most of them use a Siamese network to determine whether two input image patches contain the same forensic characteristics, such as EXIF metadata [19] or camera model characteristics [4, 30, 31]. Recently, Zhao *et al.* [63] propose to use a pair-wise similarity consistency volume [12, 64] to detect and localize Deepfakes in a fully-supervised setting. Unlike Zhao *et al.* [63] that requires a curated inconsistency image generator and pixel-level ground truth to learn the consistency volume, we only leverage image-level labels and use a self-supervised approach for training.

Weakly-supervised learning aims to use coarse or incomplete supervision to construct a model to predict fine-grained labels. For example, given image-level categorical labels to predict bounding box [21, 46] or segmentation mask [37, 38, 67] and given video-level categorical labels to predict temporal boundaries of actions [50]. Such a paradigm greatly relieves the annotation burden from its fully-supervised counterparts.

Our weakly-supervised image manipulation detection (W-IMD) is most related to weakly-supervised semantic segmentation (W-SSS), where only image-level labels can be leveraged to predict segmentation mask [37, 38, 67, 5]. Different from most W-SSS works, this paper focuses on improving the generalization ability of *image-level* manipulation detection, instead of pursuing high pixel-level localization ability. Thus, we leverage two different single-stage W-SSS methods as baselines due to their fast and simple training schemes: multiple instance learning fully convolutional network (MIL-FCN) [37] and Araslanov and Roth [1]. The former applies multiple-instance learning on weakly-supervised segmentation without considering specific prior knowledge on this task; the latter achieves strong segmentation performance by considering several priors in W-SSS, such as local consistency, semantic fidelity, and completeness.

3. Proposed Method

During training, for each input image $\boldsymbol{X} \in \mathbb{R}^{H \times W \times 3}$ with height H and width W, we only have its image-level manipulation label $y \in \{0, 1\}$, where 0 denotes authentic images, and 1 denotes manipulated images. During inference, for each image, we not only predict whether the image is tampered with, but we also generate a binary localization map $\bar{\boldsymbol{M}} \in \{0,1\}^{H\times W}$ to localize manipulation at the pixel level. An overview of our method is shown in Fig. 2.

3.1. Baseline

Without loss of generality, given an input image of size $H \times W$, we denote the final prediction map generated by a baseline method as $\hat{\boldsymbol{M}} \in (0,1)^{H\times W}$. The image-level prediction $\hat{y}$ is generated by pooling the prediction map: $\hat{y} = \text{Pool}(\hat{\boldsymbol{M}})$, where the pooling function can be method-specific, *e.g.*, global max pooling. The image classification loss $\mathcal{L}_{\text{CLS}}$ is typically a binary cross-entropy (BCE) loss between the prediction and the ground truth $\mathcal{L}_{\text{CLS}} = \mathcal{L}_{\text{BCE}}(y, \hat{y})$, where

$$\mathcal{L}_{\text{BCE}}(y, \hat{y}) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}). \tag{1}$$

The final manipulation localization map $\bar{\boldsymbol{M}}$ is obtained by thresholding $\hat{\boldsymbol{M}}$ at θ, and θ is a thresholding hyperparameter.

3.2. Adaptive Pooling for Image-Level Detection

Global max pooling has been one of the most widely used pooling methods for image-level prediction generation in image manipulation detection [42, 25, 55, 65, 6]. However, it has several clear disadvantages. First, it only detects the most discriminative part, but it fails to detect the full extent of the manipulation. Second, the loss only back-propagates through the sole maximal response, impeding the model training. To address this problem, inspired by Otsu's method of image binarization [34], we propose an adaptive pooling method, which dynamically selects pixel-level responses from the prediction map $\hat{\boldsymbol{M}}$ to form the image-level prediction $\hat{y}_{\text{A}}$.

Specifically, we first use Otsu's method to partition the prediction map into two groups. The Otsu's method finds a threshold ω_{o} that minimizes the intra-class prediction variance [34]:

$$\begin{aligned}\omega_{\text{o}} = \underset{\omega\in\{\hat{m}_{i,j}\}}{\arg\min}\, &|\{\hat{m}_{i,j}|\hat{m}_{i,j}<\omega\}|\, \text{var}\left(\{\hat{m}_{i,j}|\hat{m}_{i,j}<\omega\}\right) + \\ &|\{\hat{m}_{i,j}|\hat{m}_{i,j}\geq\omega\}|\, \text{var}\left(\{\hat{m}_{i,j}|\hat{m}_{i,j}\geq\omega\}\right),\end{aligned} \tag{2}$$

where $\text{var}(\cdot)$ denotes the variance, and $\hat{m}_{i,j}$ is the pixel-level response at spatial location (i,j) on $\hat{\boldsymbol{M}}$. As Otsu's method only applies to discrete distributions, in practice, we restrict the candidate set of thresholds ω to the values of pixel-level responses $\{\hat{m}_{i,j}\}$. The image-level prediction $\hat{y}_{\text{A}}$ determined by our adaptive pooling is the average value of the group with a higher response: $\hat{y}_{\text{A}} = \frac{1}{|\mathbb{M}_{\text{h}}|}\sum_{\hat{m}\in\mathbb{M}_{\text{h}}}\hat{m}$, where $\mathbb{M}_{\text{h}} = \{\hat{m}_{i,j}|\hat{m}_{i,j}\geq\omega_{\text{o}}\}$. And our adaptive pooling-based classification loss $\mathcal{L}_{\text{A-CLS}}$ is as the BCE loss between the ground truth and the adaptive pooling prediction: $\mathcal{L}_{\text{A-CLS}} = \mathcal{L}_{\text{BCE}}(y, \hat{y}_{\text{A}})$. By exploiting multiple high responses instead of only the maximal one, our adaptive pooling is more robust to noisy high responses and able to capture a more complete manipulation region.

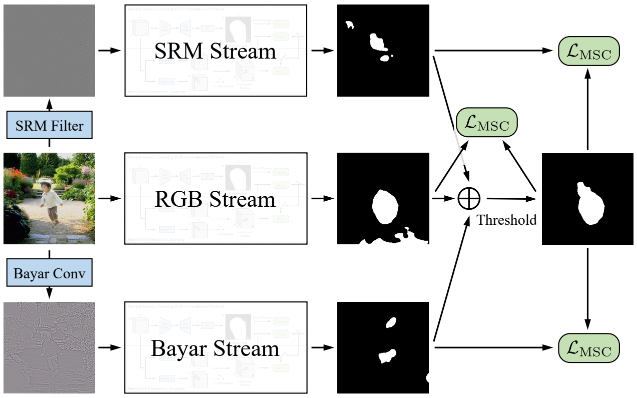

SRM Stream
RGB Stream
Bayar Stream
SRM Filter
Bayar Conv
$\mathcal{L}_{\text{MSC}}$
$\mathcal{L}_{\text{MSC}}$
$\mathcal{L}_{\text{MSC}}$
Threshold

Figure 3. An illustration of multi-source consistency learning. Three parallel streams are trained on RGB image, SRM noise map, and Bayar noise map separately. Their weighted average prediction is used as pseudo ground truth to supervise each single stream.

3.3. Learning Multi-Source Consistency

Prior arts found that exploring semantic information from images works well for IND manipulation detection, but yields poor OOD detection performance [65]. Moreover, leveraging image noise maps to learn content-agnostic information can produce strong performance [14, 66, 2, 55, 18, 6]. Given these findings, we speculate that relying exclusively on content-related information would be insufficient for detecting and localizing the manipulations. However, the success of previous research heavily relies on the pixel-level ground truth. Without such strong supervision, the model easily overfits without fully learning from different input sources. To mitigate this problem, we propose multi-source consistency (MSC) learning. First, MSC employs a multi-stream framework, with each stream exploiting different sources of the image, and thus fully exploits the manipulation in different views. By combining the outputs of individual streams, the detection and localization results can be more robust and accurate. Moreover, the ensemble prediction is used to supervise each individual stream, while helping to alleviate overfitting, correct failures within single streams, and ultimately improve the final prediction. An illustration is shown in Fig. 3.

Specifically, we adopt a three-stream framework, with each stream taking as input raw RGB image, SRM noise map [14, 66], and Bayar noise map [2], respectively. The SRM and Bayar noise maps are selected due to its wide use in previous works [66, 25, 55, 18, 57, 6, 9]. The three streams do not share parameters, but share the same training scheme. Given predictions from three streams, an ensemble prediction $\hat{\boldsymbol{M}}_{\text{ens}}$ can be obtained by weighted averaging three prediction maps:

$$\hat{\boldsymbol{M}}_{\text{ens}} = \frac{w_{\text{R}}\hat{\boldsymbol{M}}_{\text{R}} + w_{\text{S}}\hat{\boldsymbol{M}}_{\text{S}} + w_{\text{B}}\hat{\boldsymbol{M}}_{\text{B}}}{w_{\text{R}} + w_{\text{S}} + w_{\text{B}}}, \tag{3}$$

where w_{R}, $\hat{\boldsymbol{M}}_{\text{R}}$, w_{S}, $\hat{\boldsymbol{M}}_{\text{S}}$, w_{B}, $\hat{\boldsymbol{M}}_{\text{B}}$ are predefined weights and prediction maps of the RGB stream, SRM stream, and Bayar stream, respectively. Then, a binary pseudo ground truth $\bar{\boldsymbol{M}}_{\text{ens}}$ is generated by thresholding the weighted average prediction map at θ, and provides pixel-level supervision of the three streams.

Our MSC learning encourages each individual stream to fit the pseudo ground truth with a BCE loss:

$$\mathcal{L}_{\text{MSC}} = \frac{1}{HW} \sum_{i,j} \mathcal{L}_{\text{BCE}}(\bar{m}_{\text{ens},i,j}, \hat{m}_{\text{source},i,j}), \tag{4}$$

where source $\in \{\text{R}, \text{S}, \text{B}\}$. All individual streams are trained in parallel and simultaneously. Intuitively, pixels that all three streams have high activations are more likely to contain manipulations, while pixels that only one stream has high activations are less likely to be manipulated. Thus, our MSC enables each stream to improve itself by learning the voting ensemble, and in turn improves the pseudo ground truth.

3.4. Learning Inter-Patch Consistency

Only exploiting local image features fails when the tampered region is larger than the final layer's receptive field, as tampered regions in manipulations like copy-move and splicing are both from authentic images, and they both have unified, authentic forensic characteristics. Without referring to the global context, it is intractable to detect these manipulations. Thus, we propose to learn global inter-patch similarity (IPC) to detect inconsistent image patches, and differentiate low-level features between authentic and tampered patches.

IPC is conducted at an intermediate feature map $\boldsymbol{F} \in \mathbb{R}^{H' \times W' \times C}$, where H', W', and C are respectively height, width, and the number of channels. Each feature vector $\boldsymbol{f}_{i,j}$ at spatial location i, j represents a local image patch within its receptive field [27, 63]. For each patch, we compute its dot product similarity against all image patches, thus, a consistency volume $\hat{\mathbf{V}} \in (0,1)^{H' \times W' \times H' \times W'}$ containing all pair-wise similarities can be obtained:

$$\hat{v}_{i,j,h,k} = 1 - \sigma\left(\frac{\phi_1(\boldsymbol{f}_{i,j}) \cdot \phi_2(\boldsymbol{f}_{h,k})}{\sqrt{C}}\right), \tag{5}$$

where ϕ_1 and ϕ_2 are two embedding heads realized by two-layer MLPs, $\hat{v}_{i,j,h,k}$ denotes the value at location (i, j, h, k) of the consistency volume, and $\sigma(\cdot)$ denotes the sigmoid function. If patches $\boldsymbol{f}_{i,j}$ and $\boldsymbol{f}_{h,k}$ share the same forensic characteristic (*i.e.*, if they are both authentic or both tampered with), then $\hat{v}_{i,j,h,k} = 0$, while $\hat{v}_{i,j,h,k} = 1$ indicates different forensic characteristics. Therefore, authentic images are expected to have all zero consistency volumes, while tampered images should contain at least one location of value 1 in their consistency volumes. Fig. 2 bottom branch shows an illustration.

Split	Dataset	#au	#tp	#cpmv	#splc	#inpaint
Train	CASIAv2 [11]	7,491	5,063	3,235	1,828	0
	GIER [43]	4,189	4,190	N/A		
	IEdit [45]	2,255	2,255	N/A		
Val	IMD2020 [33]	2,010	2,010	N/A		
Test	CASIAv1 [10]	800	920	459	461	0
	Columbia [17]	183	180	0	180	0
	Coverage [51]	100	100	100	0	0
	NIST16 [15]	0	563	68	288	208
	GIER [43]	452	618	N/A		
	IEdit [45]	401	445	N/A		

Table 1. Dataset details. "cpmv", "splc" are abbreviations for copy-move and splicing, respectively.

Despite the previous exploration where the consistency volume is trained with full supervision and a hand-crafted inconsistency image generator [63], we show that the consistency volume can be learned in a self-supervised learning way, and benefit from multi-source consistency learning.

In the fully-supervised setting, such consistency volume can be easily learned from the pixel-level ground truth [63]. However, in the weakly-supervised setting, such ground truth is unavailable. To mitigate this problem, we exploit two different ways to generate the learning target of IPC.

Self-supervision is inspired by self-distillation [62], where the intuition is that deeper layers enjoy larger receptive fields and can make better predictions compared to shallow layers. In this setting, we regard the final localization map $\bar{\boldsymbol{M}}$ as the teacher to guide the consistency volume learning within each individual stream.

Ensemble-supervision uses the MSC ensemble localization map $\bar{\boldsymbol{M}}_{\text{ens}}$ as the IPC learning target. Ensemble supervision fuses predictions from multiple sources and can help individual streams learn cross-source information.

Given the target localization map $\bar{\boldsymbol{M}}_{\text{tgt}} \in \{\bar{\boldsymbol{M}}, \bar{\boldsymbol{M}}_{\text{ens}}\}$, we first downsample it to size (H', W'), then convert it to the target consistency volume $\bar{\mathbf{V}}_{\text{tgt}}$: if location (i, j) and (h, k) share the same value in the downsampled map, then $\bar{v}_{\text{tgt},i,j,h,k} = 0$, otherwise, $\bar{v}_{\text{tgt},i,j,h,k} = 1$. The IPC is supervised by the BCE loss:

$$\mathcal{L}_{\text{IPC}} = \frac{1}{H'W'H'W'} \sum_{i,j,h,k} \mathcal{L}_{\text{BCE}}(\bar{v}_{\text{tgt},i,j,h,k}, \hat{v}_{i,j,h,k}). \tag{6}$$

In this way, our IPC enhances the low-level feature representation, differentiates authentic and tampered image patches in shallow layers and in turn boosts the final prediction.

3.5. Optimization and Inference

The overall loss $\mathcal{L}_{\text{total}}$ is a weighted sum of the adaptive pooling-based image classification loss $\mathcal{L}_{\text{A-CLS}}$, the MSC

	Method	CASIAv1				Columbia				Coverage				IMD2020				Avg	
		AUC	Spe.	Sen.	I-F1	AUC	Spe.	Sen.	I-F1	AUC	Spe.	Sen.	I-F1	AUC	Spe.	Sen.	I-F1	AUC	I-F1
Un.	NOI1 [29]	0.500	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000
	CFA1 [13]	0.482	0.000	1.000	0.000	0.344	0.000	1.000	0.000	0.525	0.000	1.000	0.000	0.500	0.000	1.000	0.000	0.500	0.000
Full	Mantra-Net [55]	0.141	0.000	1.000	0.000	0.701	0.000	1.000	0.000	0.491	0.000	1.000	0.000	0.719	0.000	1.000	0.000	0.513	0.000
	CR-CNN [57]	0.766	0.224	0.930	0.361	0.783	0.246	0.961	0.392	0.566	0.070	0.967	0.131	0.617	0.112	0.936	0.200	0.683	0.271
	GSR-Net [65]	0.502	0.011	0.994	0.022	0.502	0.011	1.000	0.022	0.515	0.000	1.000	0.000	0.505	0.008	0.998	0.014	0.506	0.019
	CAT-Net [22]	0.630	0.328	0.762	0.459	0.849	0.373	0.782	0.505	0.572	0.093	0.902	0.169	0.721	0.132	0.872	0.229	0.693	0.157
	FCN+DA [6]	0.796	0.844	0.717	**0.775**	0.762	0.322	0.950	0.481	0.541	0.100	0.900	0.180	**0.746**	0.100	0.981	0.182	0.711	0.404
Weak	MIL-FCN [37]	0.647	0.538	0.569	0.553	0.807	0.220	0.732	0.338	0.542	0.062	0.793	0.115	0.578	0.116	0.886	0.205	0.644	0.303
	MIL-FCN [37] + WSCL	**0.829**	0.795	0.690	0.738	**0.920**	0.519	0.983	**0.680**	0.584	0.440	0.714	**0.544**	0.733	0.221	0.966	**0.360**	**0.766**	**0.580**
	Araslanov and Roth [1]	0.642	0.458	0.542	0.496	0.773	0.127	0.902	0.223	0.560	0.077	0.746	0.140	0.665	0.126	0.832	0.219	0.660	0.270
	Araslanov and Roth [1] + WSCL	0.796	0.638	0.726	0.679	0.917	0.324	0.948	0.483	**0.591**	0.220	0.838	0.348	0.701	0.193	0.872	0.316	0.751	0.456

Table 2. Comparison with unsupervised manipulation localization methods and fully-supervised methods on image-level manipulation detection. The best and the second best results are noted with **boldface** and underlined, respectively.

	Method	GIER [43]		IEdit [45]		Avg	
		AUC	F1	AUC	F1	AUC	I-F1
Full	CAT-Net [22]	0.508	0.336	0.532	0.476	0.502	0.406
	FCN+DA [6]	0.507	0.428	0.539	0.489	0.523	0.458
	MVSS-Net [6]	0.510	0.325	0.537	0.522	0.523	0.423
Weak	MIL-FCN [37] + WSCL	0.574	0.320	0.563	0.556	0.568	0.438
	MIL-FCN [37] + WSCL w/ fine-tune	**0.621**	**0.533**	**0.617**	**0.602**	**0.619**	**0.568**

Table 3. Image-level manipulation detection performance comparison on novel image manipulation datasets [43, 45].

loss $\mathcal{L}_{\text{MSC}}$, and the IPC loss $\mathcal{L}_{\text{IPC}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{A-CLS}} + w(t)\lambda_{\text{MSC}}\mathcal{L}_{\text{MSC}} + w(t)\lambda_{\text{IPC}}\mathcal{L}_{\text{IPC}}, \quad (7)$$

where λ_{MSC} and λ_{IPC} are weighting hyperparameters. $w(t) = \exp(-5(1 - t/T)^2)$ is a time-dependent Gaussian warming-up function, where t is the current epoch, and T is the maximal number of epochs. Such a warming-up function prevents learning from unreliable pseudo ground truth at early training stages [47, 23].

The final image-level prediction is obtained by weighted averaging the predictions from three different streams, and the prediction map is the ensemble localization map $\bar{M}_{\text{ens}}$.

4. Experiments

Datasets. Without explicitly mentioning, we train our method on CASIAv2 [11] only. For the in-distribution (IND) evaluation, we use the CASIAv1 dataset [10]. For the out-of-distribution (OOD) evaluation, we use three datasets: Columbia [17], Coverage [51] and IMD2020 [33]. We further follow the convention to use NIST16 [15] for the pixel-level manipulation localization evaluation. We use the IMD2020 dataset [33] for validation and hyperparameter tuning. IMD2020 contains $2,010$ real-life manipulated images, and we randomly sample $2,010$ images from the real image set as the authentic image set. To demonstrate the capacity of our method for novel manipulations, we carry out experiments on recent language-driven image editing datasets GIER [43] and IEdit [45]. Both datasets consist of paired images before and after editing, and the manipulations are not constrained by copy-move, splicing, and inpainting. To avoid data leakage from the paired images in the two datasets, we only sample either an authentic image or an edited image from each pair to form the training set. Details on the datasets are listed in Tab. 1.

Evaluation metrics. For image-level manipulation detection, we report specificity, sensitivity, and their F1 score (I-F1). The area under receiver operating characteristic (AUC) is also reported as a threshold-agnostic metric for image-level detection. For pixel-level manipulation localization, we follow previous methods [66, 42, 65, 6] to compute pixel-level precision, recall, and their F1 (P-F1) score on tempered images. The overall performance of image- and pixel-level manipulation detection/localization is measured by the harmonic mean of pixel-level and image-level F1 scores [6], denoted as combined F1 (C-F1), and is sensitive to the lower value of P-F1 and I-F1. To ensure a fair comparison, a decision threshold of 0.5 is employed for all methods when performing F1 computations.

Implementation details. Our method is implemented with PyTorch [35]. We use ResNet50 [16] as the backbone, and its weight is randomly initialized. The input size $H \times W$ is set to 224×224. Only random cropping and flipping are used for data augmentation. We use the AdamW optimizer [28] with a learning rate that decays from 10^{-4} to 10^{-5} and a weight decay factor 5×10^{-4}. We train the model for 60 epochs, and the learning rate decays by a factor of 0.1 at the 50-th epoch. Following [6], we use $\theta = 0.5$ as the default threshold to conduct the experiments. The hyperparameters are determined via a grid search on the validation set: $\lambda_{\text{MSC}} = 0.1$, $\lambda_{\text{IPC}} = 0.1$, and $w_{\text{R}} : w_{\text{S}} : w_{\text{B}} = 1 : 2 : 2$. The consistency volume is computed after the first residual block in ResNet50, and its size is $H' \times W' = 56 \times 56$.

4.1. Comparison with the State-of-the-art

As our essential goal is to improve the generalization ability for *image-level* manipulation detection, we build our WSCL upon two single-stage W-SSS methods for their fast and simple training schemes: MIL-FCN [37] and Araslanov and Roth [1]. The former is a multiple instance learning based method without considering priors in semantic information; the latter is a state-of-the-art single-stage W-SSS method with several domain-specific designs.

Image-level manipulation detection results are shown in Tab. 2. Our WSCL improves both MIL-FCN [37] and

	Method	Pixel-Level F1						Combined F1				
		CASIAv1	Columbia	Coverage	IMD2020	NIST16	Avg	CASIAv1	Columbia	Coverage	IMD2020	Avg
Un.	NOI1 [29]	0.157	0.311	0.205	0.124	0.089	0.190	0.000	0.000	0.000	0.000	0.000
	CFA1 [13]	0.140	0.320	0.188	0.111	0.106	0.188	0.000	0.000	0.000	0.000	0.000
Full	Mantra-Net [55]	0.155	0.364	0.286	0.122	0.000	0.185	0.000	0.000	0.000	0.000	0.000
	CR-CNN [57]	0.405	0.436	0.291	-	0.238	-	0.382	0.413	0.181	-	-
	GSR-Net [65]	0.387	0.613	0.285	0.175	0.283	0.349	0.042	0.042	0.000	0.026	0.028
	CAT-Net [22]	0.276	0.352	0.134	0.102	0.138	0.200	0.345	0.406	0.149	0.144	0.261
	FCN+DA [6]	0.441	0.223	0.199	0.270	0.167	0.260	0.562	0.305	0.189	0.217	0.318
Weak	MIL-FCN [37]	0.117	0.089	0.121	0.097	0.024	0.090	0.193	0.141	0.118	0.131	0.146
	MIL-FCN [37] + WSCL	0.172	0.270	0.178	0.193	0.110	0.185	0.280	0.386	0.268	0.252	0.296
	Araslanov and Roth [1]	0.112	0.102	0.127	0.094	0.026	0.092	0.182	0.140	0.133	0.046	0.125
	Araslanov and Roth [1] + WSCL	0.153	0.362	0.201	0.173	0.099	0.198	0.250	0.414	0.255	0.159	0.270

Table 4. Comparison with unsupervised and fully-supervised methods on pixel-level manipulation localization and the combined F1 score between I-F1 and P-F1. The pixel-level manipulation localization performances are measured on manipulated images only.

Araslanov and Roth [1] baselines on all datasets. Our WSCL with both baselines compares favorably with the previous fully-supervised methods in terms of detection AUC and F1. We note that our WSCL with Araslanov and Roth baseline [1] underperforms that with the MIL-FCN baseline [37]. Such results indicate that the priors in W-SSS (*e.g.*, local consistency and semantic fidelity) may not help in W-IMD, and developing specific methods for W-IMD is imperative. We evaluate the IND and OOD manipulation detection with the MIL-FCN baseline [37] in Fig. 1, where we observe a strong OOD manipulation detection performance of our method that surpasses previous fully-supervised methods, showing the effectiveness of our WSCL. Note that for unsupervised methods [29, 13], we use the maximal response on the prediction map as its image-level prediction. They tend to detect manipulations in all images, resulting in near 0.5 AUC and 0.0 F1 scores.

Novel manipulation detection. As emerging novice-friendly manipulation methods [44, 20, 44, 39, 59, 58, 61] do not necessarily generate pixel-level masks during their editing process, existing fully-supervised methods cannot make use of these weakly-labeled data. We investigate the capacity of fully-supervised methods and our weakly-supervised method on two additional datasets [43, 45], which contain manipulations that are different from the standard setting (*i.e.*, copy-move, splicing, and inpainting). The results are summarized in Tab. 3, where MIL-FCN is used as baseline [37] due to its strong image-level manipulation detection performance observed in Tab. 2. Without fine-tuning, our method already outperforms fully-supervised counterparts at the average AUC on both datasets. Such results demonstrate a strong generalization ability of our WSCL. We further fine-tune our model with image-level labels on the two datasets, and achieve the best performance. Though the comparison between our fine-tuned model and fully-supervised methods may not be fair, they cannot be trained without pixel-level mask, demonstrating the necessity of developing weakly-supervised methods.

Pixel-level manipulation localization results are listed in the left part of Tab. 4. Our method achieves reasonable pixel-level manipulation localization performance, and the average performance on five datasets is comparable with fully-supervised Mantra-Net [55] and CAT-Net [22]. Such a strong performance demonstrates the capability of our pixel-level manipulation localization.

Method	Image-Level				P-F1	C-F1
	AUC	Spe.	Sen.	I-F1		
Max Pool	0.578	0.116	0.886	0.205	**0.131**	0.131
Avg Pool	0.569	0.076	0.902	0.140	0.082	0.103
GeM [40]	0.683	0.139	0.725	0.233	0.111	0.149
GSM [49]	0.679	0.105	0.833	0.186	0.127	0.151
AP (Ours)	**0.693**	0.162	0.788	**0.269**	0.116	**0.162**

Table 5. Comparison on different pooling methods on the IMD2020 [33] dataset with RGB as the source.

Overall detection and localization performance is summarized in the right part of Tab. 4. Our method achieves a similar average performance with CAT-Net [22]. Surprisingly, our method achieves the best overall performance on the Coverage dataset [51]. Such a strong performance demonstrates the effectiveness of our method.

Qualitative results are visualized in Fig. 4. We make the following observations. (1) Predictions from unsupervised methods [29, 13] tend to be noisy, while both fully-supervised method and our weakly-supervised method generate clean localization maps. (2) Our method tends to predict a large area of manipulation, encompassing the ground truth area, while fully-supervised method detects clean manipulation boundaries. (3) Our method shows degraded results on copy-move (see Coverage examples), where both source and target manipulated areas are from the same image, and our method tends to detect both areas.

4.2. Ablation Study

We carry out the ablation study on an OOD dataset IMD2020 [33] with MIL-FCN [37] as the baseline, where all variants are trained with the CASIAv2 dataset [11]. The ablation study is performed progressively, with each subsequent setting using the previous one as a baseline.

Adaptive pooling. To mitigate the problem of max pooling in previous methods, we propose an adaptive pooling

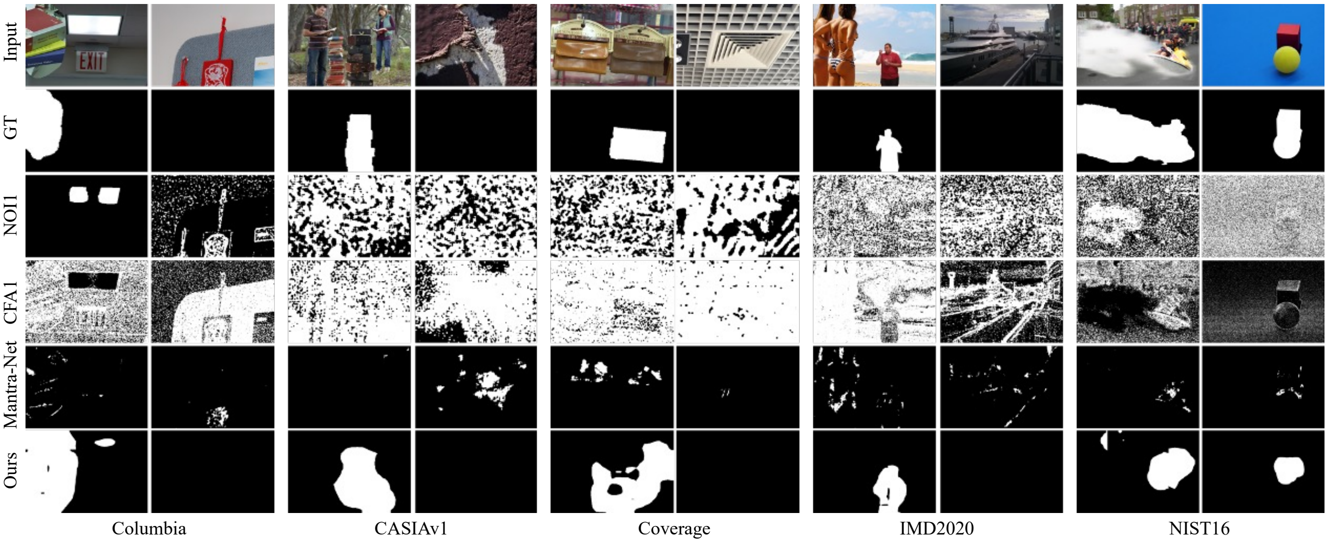


Figure 4. Qualitative results on five datasets. From top to bottom in each group: input image, ground truth mask, and predictions from NOI1 [29], CFA1 [13], Mantra-Net [55] and our WSCL with MIL-FCN [37] as the baseline.

	Source	Image-Level				P-F1	C-F1
		AUC	Spe.	Sen.	F1		
w/o MSC	RGB	0.693	0.162	0.788	0.269	0.116	0.162
	Bayar	0.685	0.187	0.642	0.290	0.132	0.181
	SRM	0.674	0.196	0.733	0.309	0.109	0.161
	Fusion	0.701	0.209	0.748	0.327	0.143	0.199
w/ MSC	RGB	0.701	0.185	0.836	0.303	0.136	0.188
	Bayar	0.715	0.193	0.762	0.308	0.177	0.225
	SRM	0.707	0.210	0.837	0.336	0.181	0.235
	Fusion	**0.726**	0.218	0.857	**0.348**	**0.188**	**0.244**

Table 6. Ablation study on the multi-source consistency learning on IMD2020 [33].

IPC	Image-Level				P-F1	C-F1
	AUC	Spe.	Sen.	F1		
w/o	0.726	0.218	0.857	0.348	0.188	0.244
self.	0.730	0.219	0.920	0.354	0.192	0.249
ens.	**0.733**	0.221	0.966	**0.360**	**0.193**	**0.252**

Table 7. Ablation study on the inter-patch consistency learning on IMD2020 [33].

to dynamically assign image-level labels to the pixels. We compare our adaptive pooling with related pooling methods [49, 40] in Tab. 5. The results show adaptive pooling achieves the best performance on all major metrics. Besides, both GSM and GeM introduce an additional hyperparameter, while our adaptive pooling does not require any hyperparameter, making it more flexible. Such advantage demonstrates the effectiveness of our adaptive pooling.

Multi-source consistency learning promotes unanimous predictions among all individual models through a pixel-level pseudo ground truth. The results are summarized in Tab. 7. We observe that the late fusion improves the single-stream performance w/ and w/o MSC, showing the effect of voting ensemble. Furthermore, our MSC improves the performance on single streams and the fusion results, demonstrating its effectiveness on improving generalization.

Inter-patch consistency learning aims to learn global patch-patch similarities, and further differentiate low-level authentic and tampered image patch features. Two different supervision implementations are tested: the localization map from the same stream (self-supervision), and the ensemble localization map from three streams (ensemble-supervision). We make the following observations from Tab. 7. (1) Both implementations of IPC clearly improve overall performance, showing the effectiveness of IPC. (2) The ensemble-supervision IPC outperforms the self-supervision counterpart, and this is intuitive as the ensemble target fuses information from multiple sources.

5. Conclusion

We propose the task of weakly-supervised image manipulation detection, such that only binary image-level labels are required to detect and localize manipulations. We propose a weakly-supervised self-consistency learning for this task that aims to improve the generalization ability. Two different self-consistency learning schemes are employed: multi-source consistency and inter-patch consistency. By leveraging content-agnostic information and combining predictions from various sources, MSC enhances the individual stream's ability for both manipulation detection and localization. IPC learns the global similarity between image patches to detect a complete region of manipulation, which improves the low-level representation of image patches and thus facilitates the MSC learning process. Our WSCL shows strong image-level manipulation detection performance under both IND and OOD evaluation settings. We also achieve reasonable pixel-level manipulation localization performance.

Limitations. While our method demonstrates robust performance in detecting image-level manipulation, its capability in localizing the pixel-level manipulation is merely satisfactory. This is an important limitation as accurate localization is key to explainability and understanding the extent of the manipulation, which is vital in forensics applications. Besides, as shown in our supplementary material, our method is vulnerability to certain types of noise and distortions, such as Gaussian blur, which could potentially be exploited to bypass our detection method. This highlights the need for the method to be robust not just against various manipulation techniques, but also against different types of image noise and distortions. Thus, future work should aim to improve the pixel-level manipulation localization ability, and the robustness against different image distortions.

Acknowledgements

This work is supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0124. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

References

[1] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, pages 4253–4262, 2020. 3, 6, 7

[2] 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*, pages 2691–2706, 2018. 2, 3, 4

[3] Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Improved dct coefficient analysis for forgery localization in jpeg images. In *ICASSP*, pages 2444–2447, 2011. 2

[4] Luca Bondi, Luca Baroffio, David Güera, Paolo Bestagini, Edward J Delp, and Stefano Tubaro. First steps toward camera model identification with convolutional neural networks. *IEEE Signal Processing Letters*, pages 259–263, 2016. 3

[5] Lyndon Chan, Mahdi S Hosseini, and Konstantinos N Plataniotis. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *IJCV*, pages 361–384, 2021. 3

[6] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *ICCV*, 2021. 1, 3, 4, 6, 7

[7] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security*, pages 2284–2297, 2015. 2

[8] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Splicebuster: A new blind image splicing detector. In *2015 IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2015. 2

[9] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE TPAMI*, 2022. 1, 3, 4

[10] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database 2010. `http://forensics.idealtest.org`, 2010. 1, 5, 6

[11] 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. 1, 5, 6, 7

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. 3

[13] 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*, pages 1566–1577, 2012. 2, 6, 7, 8

[14] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, pages 868–882, 2012. 2, 3, 4

[15] 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*, pages 63–72, 2019. 1, 5, 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[17] Yu-Feng Hsu and Shih-Fu Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *ICME*, pages 549–552, 2006. 1, 5, 6

[18] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *ECCV*, pages 312–328, 2020. 3, 4

[19] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *ECCV*, pages 101–117, 2018. 2, 3

[20] Wentao Jiang, Ning Xu, Jiayun Wang, Chen Gao, Jing Shi, Zhe Lin, and Si Liu. Language-guided global image editing via cross-modal cyclic mechanism. In *ICCV*, pages 2115–2124, 2021. 1, 2, 7

[21] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, pages 350–365, 2016. 3

[22] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *WACV*, pages 375–384, 2021. 1, 3, 6, 7

[23] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 6

[24] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *CVPR*, pages 7880–7889, 2020. 1

[25] Haodong Li and Jiwu Huang. Localization of deep inpainting using high-pass fully convolutional network. In *ICCV*, pages 8301–8310, 2019. 1, 3, 4

[26] Xihui Liu, Zhe Lin, Jianming Zhang, Handong Zhao, Quan Tran, Xiaogang Wang, and Hongsheng Li. Open-edit: Open-domain image manipulation with open-vocabulary instructions. In *ECCV*, pages 89–106, 2020. 1

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 5

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 6

[29] Babak Mahdian and Stanislav Saic. Using noise inconsistencies for blind image forensics. *Image and Vision Computing*, pages 1497–1503, 2009. 2, 6, 7, 8

[30] Owen Mayer and Matthew C Stamm. Learned forensic source similarity for unknown camera models. In *ICASSP*, pages 2012–2016, 2018. 3

[31] Owen Mayer and Matthew C Stamm. Forensic similarity for digital images. *IEEE Transactions on Information Forensics and Security*, pages 1331–1346, 2019. 3

[32] Owen Mayer and Matthew C Stamm. Exposing fake images with forensic similarity graphs. *IEEE Journal of Selected Topics in Signal Processing*, pages 1049–1064, 2020. 3

[33] 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 Workshops*, pages 71–80, 2020. 1, 5, 6, 7, 8

[34] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE TPAMI*, pages 62–66, 1979. 4

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035. 2019. 6

[36] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, pages 2085–2094, 2021. 1, 2

[37] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR*, 2015. 3, 6, 7, 8

[38] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, pages 1713–1721, 2015. 3

[39] Tiziano Portenier, Qiyang Hu, Attila Szabó, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: deep sketch-based face image editing. (4):1–13, 2018. 1, 2, 7

[40] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE TPAMI*, (7):1655–1668, 2018. 7, 8

[41] Yuan Rao and Jiangqun Ni. A deep learning approach to detection of splicing and copy-move forgeries in images. In *2016 IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2016. 2

[42] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. Image splicing localization using a multi-task fully convolutional network (mfcn). *Journal of Visual Communication and Image Representation*, pages 201–209, 2018. 1, 2, 4, 6

[43] Jing Shi, Ning Xu, Trung Bui, Franck Dernoncourt, Zheng Wen, and Chenliang Xu. A benchmark and baseline for language-driven image editing. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 5, 6, 7

[44] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing. In *CVPR*, pages 13590–13599, 2021. 1, 2, 7

[45] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. In *ACL*, pages 1873–1883, 2019. 5, 6, 7

[46] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE TPAMI*, pages 176–191, 2018. 3

[47] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204, 2017. 6

[48] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 1

[49] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017. 7, 8

[50] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, pages 4325–4334, 2017. 3

[51] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage—a novel database for copy-move forgery detection. In *ICIP*, pages 161–165, 2016. 1, 2, 5, 6, 7

[52] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Deep matching and validation network: An end-to-end solution to constrained image splicing localization and detection. In *ACM MM*, pages 1480–1502, 2017. 2

[53] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Buster-net: Detecting copy-move image forgery with source/target localization. In *ECCV*, pages 168–184, 2018. 2

[54] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Image copy-move forgery detection via an end-to-end deep neural network. In *WACV*, pages 1907–1915, 2018. 2

[55] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and

localization of image forgeries with anomalous features. In *CVPR*, pages 9543–9552, 2019. 1, 3, 4, 6, 7, 8

[56] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *CVPR*, pages 2256–2265, 2021. 1

[57] Chao Yang, Huizhou Li, Fangting Lin, Bin Jiang, and Hao Zhao. Constrained r-cnn: A general image manipulation detection model. In *ICME*, pages 1–6, 2020. 1, 3, 4, 6, 7

[58] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In *ECCV*, pages 601–617, 2020. 1, 2, 7

[59] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, pages 4471–4480, 2019. 1, 2, 7

[60] Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. Large-scale evaluation of splicing localization algorithms for web images. *Multimedia Tools and Applications*, pages 4801–4834, 2017. 1

[61] Yu Zeng, Zhe Lin, and Vishal M Patel. Sketchedit: Mask-free local image manipulation with partial sketches. *arXiv preprint arXiv:2111.15078*, 2021. 1, 2, 7

[62] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE TPAMI*, 2021. 5

[63] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *ICCV*, pages 15023–15033, 2021. 3, 5

[64] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Recognize actions by disentangling components of dynamics. In *CVPR*, pages 6566–6575, 2018. 3

[65] 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 *AAAI*, pages 13058–13065, 2020. 1, 3, 4, 6, 7

[66] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *CVPR*, pages 1053–1061, 2018. 3, 4, 6

[67] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *CVPR*, pages 3791–3800, 2018. 3

[68] Xinshan Zhu, Yongjun Qian, Xianfeng Zhao, Biao Sun, and Ya Sun. A deep learning approach to patch-based image inpainting forensics. *Signal Processing: Image Communication*, pages 90–99, 2018. 2