Image Manipulation Detection With Implicit Neural Representation and Limited Supervision

Zhenfei Zhang[1], Mingyang Li[2], Xin Li[1], Ming-Ching Chang[1], and Jun-Wei Hsieh[3]

[1] University at Albany, State University of New York
[2] Stanford University
[3] National Yang Ming Chiao Tung University
{zzhang45, xli48, mchang2}@albany.edu
mingyang.li@stanford.edu jwhsieh@nycu.edu.tw

Abstract. Image Manipulation Detection (IMD) is becoming increasingly important as tampering technologies advance. However, most state-of-the-art (SoTA) methods require high-quality training datasets featuring image- and pixel-level annotations. The effectiveness of these methods suffers when applied to manipulated or noisy samples that differ from the training data. To address these challenges, we present a unified framework that combines unsupervised and weakly supervised approaches for IMD. Our approach introduces a novel pre-processing stage based on a controllable fitting function from Implicit Neural Representation (INR). Additionally, we introduce a new selective pixel-level contrastive learning approach, which concentrates exclusively on high-confidence regions, thereby mitigating uncertainty associated with the absence of pixel-level labels. In weakly supervised mode, we utilize ground-truth image-level labels to guide predictions from an adaptive pooling method, facilitating comprehensive exploration of manipulation regions for image-level detection. The unsupervised model is trained using a self-distillation training method with selected high-confidence pseudo-labels obtained from the deepest layers via different sources. Extensive experiments demonstrate that our proposed method outperforms existing unsupervised and weakly supervised methods. Moreover, it competes effectively against fully supervised methods on novel manipulation detection tasks.

Keywords: Image Manipulation Detection · Implicit Neural Representation · Weakly Supervised Learning · Unsupervised Learning

1 Introduction

The emergence of diverse media tampering tools, such as Photoshop and AI editing and generation methods [10, 49, 61, 64, 70, 73], has made it increasingly convenient to manipulate media content. However, this accessibility also brings forth the concerning issue of widespread misinformation, which can precipitate serious security implications. Therefore, the development and implementation of robust tampering detection technology, namely, Image Manipulation Detection

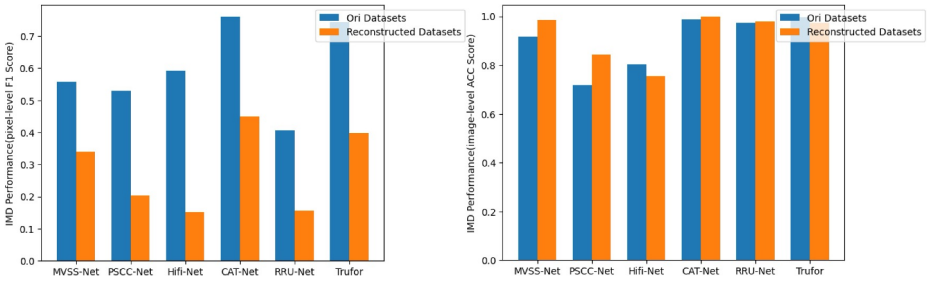

Fig. 1: We conducted experiments using three widely-used evaluation datasets containing both authentic and tampered samples. Performance are compared with six SoTA fully supervised IMD methods. The pixel-level F1 score is calculated using tampered images, while image-level accuracy is computed using authentic images. The blue and orange bars represent the original datasets and reconstructed datasets via Implicit Neural Representation, respectively. It is evident that there is a significant performance decrease in all methods when applied to reconstructed images in pixel-level detection compared with the original dataset. On the other hand, performance using authentic images shows less change. The scores are averaged across CASIAv1 [11], Coverage [57], and Columbia [24] datasets.

(IMD) methods, are imperative to mitigate these risks effectively. The fundamental manipulation operations that previous methods typically address are as follows: (1) *Splicing*, which involves taking content from one image and pasting it onto another image, (2) *Copy-move*, in which parts of an image are duplicated and relocated to another location within the same image, (3) *Inpainting*, which entails erasing parts of an image and replacing them with synthesized content.

Despite significant advances in fully supervised IMD methods, they encounter several notable challenges. First, these methods often perform poorly when confronted with unseen manipulation types. Second, extension of them towards unseen manipulation types faces challenges due to their reliance on high-quality training datasets with either image-level and pixel-level annotations. Acquiring such datasets is costly and in many cases, impractical, especially considering the myriad varieties of real-life tampering methods. Third, while some language-guided datasets may lack pixel-level labels, they hold advantages handling real-world scenarios. These datasets can potentially enhance the generalization capability of IMD models.

To address the limitations of fully supervised IMD methods and enhance the generalization ability toward real-world use, we propose to integrate unsupervised and weakly supervised approaches into a unified IMD framework. Our framework allows training with solely image-level labels or even without any labels, aligning with many unsupervised and weakly supervised tasks [14, 30, 48, 52, 56, 72, 74]. Compared to the fully supervised methods, our approach comes with superior generalization capabilities and can be trained using datasets without annotations. Our method begins with the observation that tampered regions exhibit differences from authentic regions in most cases, such as the variations

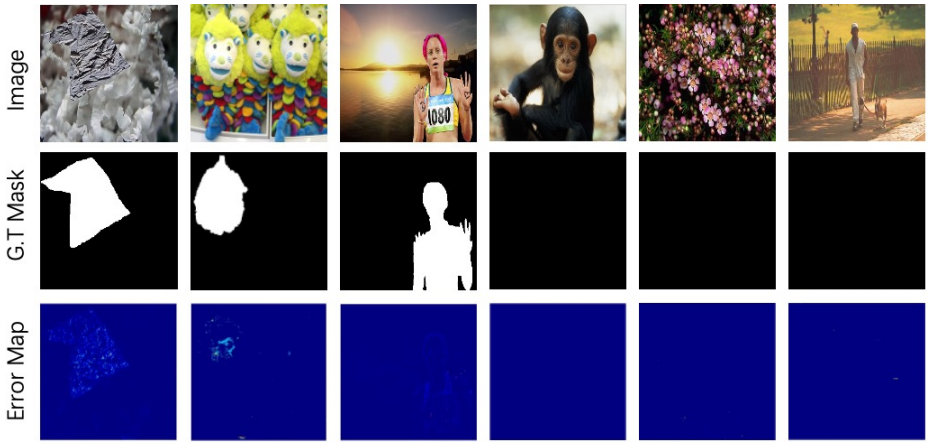

Fig. 2: Examples of Reconstruction Error Maps computed between original and reconstructed images are presented. The first two rows depict the data samples and their corresponding ground-truth masks, respectively. The first three columns showcase tampered image examples, while the last three columns display authentic images, where the ground-truth masks are all black. Apparently, the reconstruction process fails to properly reconstruct the tampered pixels, resulting in activations in the error map. Conversely, less change is observed in the authentic samples.

in color and lighting, which pose challenges for the fitting function that needs to model regions accurately. It is shown in [63] that the controllable fitting function of Implicit Neural Representation (INR) tends to learn an average representation of the training images. Motivated by this insight, we raise the following question as our hypothesis: if we train an INR solely on authentic images, can the fitting function effectively represent the characteristics of tampered regions?

To obtain the answer to this question, we first train an INR using only pristine images from CASIAv2 [12] and use it to reconstruct three mainstream datasets. We then apply fully supervised SoTA methods to evaluate the reconstructed datasets, as shown in Fig. 1. Surprisingly, the evaluation results of these methods exhibit a significant decrease when using INR-reconstructed samples, while there is less performance change in the authentic image samples. This outcome leads us to an initial assumption that the INR may not effectively capture the characteristics of tampered regions. To validate this assumption, we compute the reconstruction error map between the reconstructed and original images in Fig. 2. Remarkably, we observe activation in the tampered regions of the tampered samples, while there is no discernible difference in the authentic samples. This observation inspires us to incorporate the INR as a pre-processing method and concatenate the reconstruction error map with the input RGB images before feeding them to the backbone. We name this pre-processing method as **Neural Representation Reconstruction (NRR)**.

Following the success of the pre-processing using INR, we further explore our findings and leverage it fully in our framework. Drawing inspiration from

Contrastive Learning [22], we utilize NRR as a contrastive sample generator and introduce selective pixel-level contrastive learning, focusing solely on highly confident regions. This approach effectively mitigates uncertainty associated with the absence of pixel-level labels and further enhances weakly supervised performance. We further extend our method to a fully unsupervised approach trained with selected high-confidence pseudo-labels using a self-distillation [69] training strategy. Finally, previous SoTA methods widely apply Global-Max Pooling (GMP) or Global-Average Pooling (GAP) for image-level detection. However, GMP can hinder training and cause inaccurate predictions, as only the most discriminative response is back-propagated, neglecting the entire tampered content. Conversely, GAP is susceptible to inaccuracies due to weakly activated pixels. To overcome this limitation, we propose an adaptive global-average pooling that focuses on the high-confidence tampered regions. Our method can thus produce more comprehensive and robust image-level predictions.

Experimental evaluations are conducted on seven datasets, including five mainstream datasets featuring general manipulation types and two novel datasets containing unseen tampered samples. The results demonstrate that our methods outperform SoTA weakly and unsupervised methods. Furthermore, our method achieves competitive results compared to fully supervised methods in novel manipulation detection tasks. Finally, our method can be easily extended to the datasets without pixel-level labels, which shows enhanced generalizability.

The contribution of this paper includes the following: (1) We propose a novel method that achieves plausible weakly and unsupervised IMD results. Our method can be easily adapted to images without labels or only with image-level labels. (2) To our knowledge, we are the first to investigate the potential of Implicit Neural Representation (INR) in the IMD task. The pre-process step utilizing INR demonstrates effectiveness in handling tampered cases. (3) We introduce selective supervision, which mitigates uncertainty associated with the absence of labels and further improves detection performance. (4) Extensive experiments validate the efficacy of our proposed methods, showcasing superior performance on both standard and novel manipulation types compared to SoTA methods.

2 Related Work

2.1 Image Manipulation Detection

Most traditional unsupervised IMD methods [1, 8, 9, 17, 38] detect manipulation via low-level tampering artifacts including camera fingerprint, double compression, color filter array, *etc.* The weakly supervised method of [65] applies self-consistency learning from multiple inputs. Most fully supervised IMD [5, 20, 21, 25, 26, 33, 54, 58–60, 62] methods aim at detecting anomalous features. The two-branch architecture [29, 75] detects both image-editing and double-compression manipulation traces.

Unlike previous methods, we propose combining weakly and unsupervised IMD in a single framework. Meanwhile, most SoTA methods apply high-pass fil-

ters such as SRM [18] and Bayar [62] to suppress low-frequency information and detect noise inconsistency, but they are not adaptive to all manipulation types, especially when the tampered pixels are from the same source as authentic ones. Additionally, noise filters are very sensitive to high-frequency information such as edges, whether manipulated or not, making them ineffective and inefficient. We thus propose a novel pre-processing method using INR, which applies reconstruction error to address this limitation. Our pre-process via reconstruction error can effectively provide a good prior to the model and is non-sensitive to authentic images, as shown in Fig. 2.

2.2 Image Neural Representation

Image Neural Representation (INR) has become increasingly popular for image representation. The controllable fitting ability of INR is widely utilized in various applications such as continuous image/video super-resolution [6,7], video/image compression [13, 28], continuous shape representation [16, 66], medical image analysis [40,67], etc. Recently, [63] utilized INR for low-light image enhancement. However, the capability of INR on distinguishing pristine from tampered images has not been documented in the open literature.

2.3 Contrastive Learning

The basic idea of contrastive learning [22] is to repel negative sample pairs while attracting positive pairs. The most popular architecture using contrastive learning is the Siamese network [27], which accepts two inputs simultaneously and the process is supervised by their similarity. This method has been widely applied in unsupervised and self-supervised applications [4, 36, 53, 71]. Traditional unsupervised contrastive learning approaches commonly apply simple image augmentation operations, which may not be easily applicable to IMD due to model uncertainty [26]. Furthermore, image-level similarity, which is commonly used in contrastive learning, cannot be directly applied to IMD as it is unrelated to objects. To apply contrastive learning in un-/weakly supervised IMD, we utilize INR-reconstructed images as counter samples and employ a feature matching process in the final feature space. Building on the potential ability of INR mentioned in Section 1, authentic features tend to have higher matching scores, while tampered features have lower matching scores.

3 Proposed Method

3.1 Overall Architecture

Fig. 3 illustrates the overall architecture of our IMD framework. The basic architecture comprises two branches with shared weights. Given an RGB image $I \in \mathbb{R}^{H \times W \times 3}$ where H and W are its height and width, respectively, we first apply Neural Representation Reconstruction (NRR) to reconstruct it as

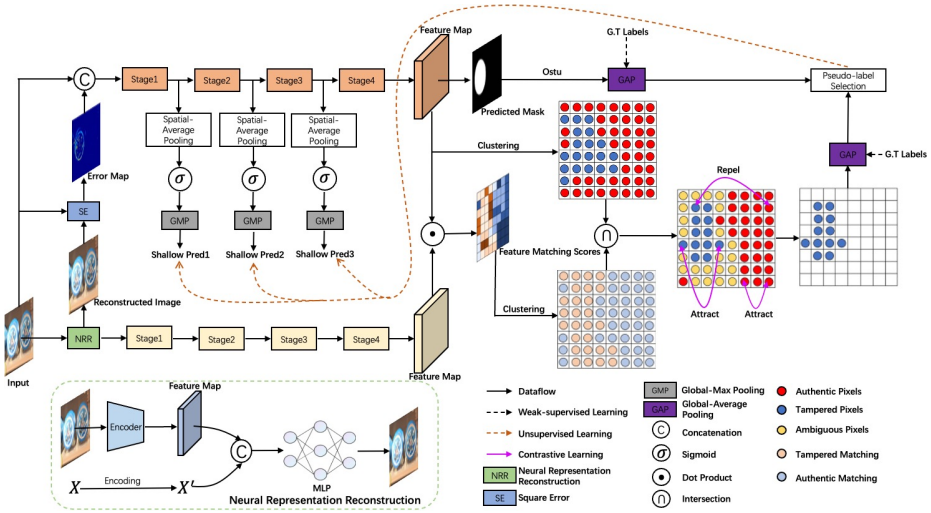

Fig. 3: An overview of the proposed two-branch framework. The first branch accepts concatenated inputs as the main branch, while the NRR reconstructed image is fed into the second branch as a complementary branch. Selective contrastive learning is applied only to the pixels that have high confidence of being authentic or tampered. The classification results conducted by global-average pooling on both the result of the main branch using Otsu's method and intersected tampered pixels from clustering are used for loss computing. In the weakly supervised setting, ground-truth image-level labels are applied for supervision. In the unsupervised setting, high-confidence pseudo-labels from the deepest layers are used to guide the shallow outputs.

$I_R \in \mathbb{R}^{H\times W\times 3}$ and generate a reconstruction error map $I_E \in \mathbb{R}^{H\times W\times 1}$ between I_R and I. We then concatenate I and I_E, feeding them into the first branch, which serves as the main branch. Similar to most IMD methods, the main branch generates a mask using a simple upsampling and Sigmoid activation function on the final feature map. We then apply Otsu's method to adaptively select the activated region for image-level prediction, as done in [65]. The reconstructed image I_R is fed into the second branch, acting as a complementary branch for feature matching. After processing with the backbone, we obtain two feature maps F and F_R. We next compute the feature matching scores M between the two feature maps via a dot product, where authentic pixels tend to have higher matching scores and vice versa. For the two-class classification of manipulation detection, unsupervised clustering is applied to F and M. We then intersect the two clustering results and exclusively apply pixel-level contrastive learning to the intersected features that exhibit higher confidence in being either authentic or tampered. The image-level classification result is conducted using the proposed adaptive global average pooling, which focuses on the high-confidence tampered regions for comprehensice image-level prediction. In a weakly supervised manner, the ground-truth image-level label is applied to supervise the prediction. In an unsupervised manner, a selected set of high-confidence pseudo-

labels from the deepest layer are utilized to supervise the shallow prediction via self-distillation [69] training strategy. The high-confidence pseudo-labels are chosen by comparing predictions derived from Otsu's method and clustering technique, opting solely for those consistently identified by both sources.

3.2 Neural Representation Reconstruction

Inspired by [63] and the observation from our experiment in Fig. 1 2, we apply NRR to reconstruct the input image. The reconstruction error can highlight the manipulation trace, thereby furnishing an indispensable prior to the subsequent IMD model. In INR, the input image is first converted to a feature map $F_N \in \mathbb{R}^{H\times W\times C}$ using an image encoder, where H and W are height and width, C is the number of feature channels. The coordinate set of input can be expressed using $X \in \mathbb{R}^{H\times W\times 2}$. We proceed by concatenating F_N and X, subsequently feeding them into a Multi-Layer Perceptron (MLP) for decoding. The NRR is formulated as:

$$I_R[x, y] = MLP(F_N[x, y], X[x, y]), \tag{1}$$

where I_R is reconstructed RGB pixel values from I, $[x, y]$ is each pixel location. The main goal of NRR is to reconstruct the RGB values of I, with loss function formulated as:

$$\mathcal{L}_{NRR} = \|I - I_R\|_1 . \tag{2}$$

Note that such reconstruction can not depict the high-frequency pixels properly. We therefore apply the positional encoding from [39] to map X to a higher-dimensional space. Such positional encoding is expressed as:

$$X^{'} = \left(\sin(2^0\pi X), \cos(2^0\pi X), \cdots, \sin(2^{L-1}\pi X), \cos(2^{L-1}\pi X)\right), \tag{3}$$

where L is a pre-setting constant to control the fitting ability of NRR. Normally, larger L results in a more accurate fitting. In our task, we aim to avoid outputs from NRR that mirror the inputs; instead, we desire NRR to faithfully preserve information in normal (authentic) content while introducing unfaithfulness in extreme (tampered) pixels. We empirically choose $L = 8$ as the optimal trade-off.

3.3 Selective Contrastive Learning

After obtaining I_R from NRR, we calculate the reconstruction error map between I and I_R using $I_E = (I_R - I)^2$. Then, we concatenate I_E and I, enhancing the input to the backbone's first (main) branch. For the input of the second (complementary) branch, we send I_R for feature matching. We employ ResNet50 [23] as a backbone, which consists of four stages that match previous weakly supervised methods. The weights of two branches are shared. After processing by the backbone, we obtain 2 feature outputs F and F_R from different input sources. We then compute the feature matching scores M using the dot product as:

$$M_{x,y} = \sigma\left(\frac{P(F_R^{x,y}) \cdot P(F^{x,y})}{\sqrt{C}}\right), \tag{4}$$

where $M_{x,y}$ is the similarity score at the spatial location (x, y). The project head $P(\cdot)$ contains 2 convolutional layers and ReLU activation. The $\sigma(\cdot)$ denotes the sigmoid activation function, and $\sqrt{C}$ provides normalization.

Due to the ability of NRR to properly reproduce only authentic pixels (and not tampered ones), the high matching scores in M tend to correspond to the authentic parts of the image. In contrast, low scores tend to correspond to manipulated regions of the image. Due to the lack of ground-truth masks to supervise the final features, we apply unsupervised clustering for forged/pristine classification similar to [3, 37, 41, 44, 47, 58] and assume that the cluster with fewer elements is the tampered cluster. This assumption aligns with the real-world situation of current manipulation datasets. The reason is that, in most cases, the tampered region is usually much smaller than the authentic ones.

Ideally, we can apply pixel-level contrastive learning through InfoNCE [22] on M and F as [58]. However, we found that this method does not work well in our experiments, as clustering may come with low confidence due to the lack of ground-truth masks. To address this issue, we intersect on the clustering results of M and F, and denote the intersected clustering as C_I. After the intersection, we will have 2 clusters with higher confidence in being either authentic or tampered with, since they come from the same prediction from 2 different sources. We thus apply InfoNCE only to intersected pixels for contrastive learning, leaving the ambiguous pixels unchanged. This selective contrastive learning loss is formulated as:

$$\mathcal{L}_{SCL} = -\log \frac{\frac{1}{J}\sum_{j\in[1,J]} \exp(q \cdot k_j^+/\tau)}{\sum_{i\in[1,K]} \exp(q \cdot k_i^-/\tau)}, \tag{5}$$

where q is an encoded query; J and K are the number of selected positive and negative keys, respectively; τ is a temperature hyper-parameter. We set positive keys k_j^+ as pixels associated with pristine regions, whereas negative keys k_i^- correspond to pixels linked to tampered regions.

3.4 Adaptive Global Average Pooling

Many existing methods use Global-Max Pooling (GMP) and Global-Average Pooling (GAP) for image-level prediction to determine if the input is authentic or tampered. However, GMP can hinder training and cause inaccurate predictions as only the most discriminative response is back-propagated, neglecting the entire tampered content. Global-Average Pooling (GAP) is susceptible to inaccuracies due to weakly activated pixels.

To tackle these challenges, we introduce **Adaptive Global Average Pooling (AGAP)**, which focuses on the high-confidence tampered regions for comprehensive image-level prediction. Leveraging the intersection of two clustering results (discussed in Section 3.3), we initially apply Global Average Pooling (GAP) exclusively to intersected tampered regions from a clustering perspective. However, relying solely on unsupervised clustering may not guarantee optimal performance and robustness across all input types without ground-truth labels. As discussed in [32], Otsu's method performs well when the image histogram

exhibits a bimodal distribution, whereas clustering provides flexibility and the ability to handle more complex histograms. Therefore, we combine Otsu and clustering to enhance image-level prediction and training robustness. Specifically, GAP is applied to the tampered responses from both Otsu and intersected clustering results for loss computation with image-level labels. Further details on Otsu's method and clustering can be found in their respective papers [15,43].

3.5 Weakly-supervised and Unsupervised IMD

In the *weakly-supervised* IMD setting, we utilize ground truth image-level labels to supervise the prediction training using a binary cross-entropy (BCE) loss, which is:

$$\mathcal{L}_{BCE}(g, \hat{g}) = -(1-g)\log(1-\hat{g}) - g\log(\hat{g}), \tag{6}$$

where g and $\hat{g}$ are the ground-truth and prediction scores, respectively. The final classification loss in a weakly supervised manner is the sum of two BCE losses, comparing two pooling results with g.

In the *unsupervised* IMD setting, where no labels are used, we employ a self-distillation training strategy [69], using pseudo-labels from the deepest layers as a teacher to supervise the shallow outputs.

To streamline prediction results from shallow layers and mitigate computational overhead, the classification head following each middle stage of the backbone uses spatial-average pooling in the channel dimension, reshaping it into a one-channel feature map. This is followed by a sigmoid function and global-max pooling. In traditional self-distillation methods, combining ground truth loss and self-distillation enhances overall performance, but this approach is not applicable in an unsupervised context. Our experiments revealed that relying solely on self-distillation did not yield satisfactory results, as the outputs from the deepest layers may lack accuracy, hindering the training process and overall performance.

Drawing inspiration from the selective-supervised method [31], proven effective in handling noisy label datasets, we leverage its concept of selecting training examples based on the alignment between feature representation and given labels. However, in our unsupervised setting, the absence of labels poses a challenge. To overcome this hurdle, we compare predictions obtained from Otsu and clustering methods, choosing only those consistently predicted by both sources as pseudo-labels for self-distillation training.

In pseudo-label selection, predictions exceeding 0.5 are considered as tampered samples. Similar to weakly supervised setting, we employ BCE loss between selected pseudo-labels and shallow predictions for supervision. During inference, all classification heads in shallow layers are excluded to avoid unnecessary parameters.

Training objective. We first apply trained NRR through $\mathcal{L}_{NRR}$ as a pre-trained model, with all its weights frozen during IMD training. For simplicity, we use the symbol $\mathcal{L}_{cls}$ to denote the loss functions for classification in both

unsupervised and weakly supervised approaches, albeit with slight differences as described above.

The total loss for our proposed IMD, denoted as $\mathcal{L}_{total}$, is a weighted sum of both classification losses using BCE loss and the selective pixel-level contrastive learning loss:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{cls} + \beta\mathcal{L}_{SCL}, \tag{7}$$

where α and β are weighting hyperparameters.

4 Experiments

Dataset: Our model is trained using CASIAv2 [12] exclusively, which comprises 7,491 authentic samples and 5,063 tampered images. For the evaluations of the standard IMD task, we employ widely-used benchmarks, including CASIAv1 [11], Coverage [57], Columbia [24], IMD2020 [42], and NIST16 [19]. CASIAv1 [11] consists of both splicing and copy-move images. Coverage [57] contains only copy-move samples with some post-processing approaches. Columbia [24] comprises 363 uncompressed images with an average resolution of 938×720. NIST16 [19] and IMD2020 [42] contain only tampered images, suitable for pixel-level evaluation. These datasets cover traditional manipulation types including splicing, copy-move, and removal. For evaluations involving novel or more complex manipulation types, we utilize IEdit [51] and MagicBrush [68], which are two language-driven datasets containing various novel manipulation types, such as action change and light change.

Evaluation Metrics: We utilize IOU and F1 scores, including P-F1 for pixel-level F1, I-F1 for image-level F1, and C-F1 for combined F1. The C-F1 score accounts for both pixel-level and image-level performance through the harmonic mean, providing an overall performance comparison. All F1 scores and IOU scores are computed using 0.5 as the fixed threshold. Due to the lack of pixel-level masks in IEdit [51], we include image-level ACC for additional evaluation.

Implementation Details: We employ ResNet50 [23] as the backbone and the model is implemented using PyTorch [45], with parameters initialized randomly. We apply AdamW [35] as the optimizer. The Multi-Layer Perceptron (MLP) in NRR follows a three-hidden-layer architecture. NRR is trained for 120 epochs with an initial learning rate of 2×10^{-4} and weight decay is applied. The IMD model in weakly supervised mode is trained for 50 epochs with an initial learning rate of 0.0005 and weight decay. For the unsupervised model, we train for 20 epochs with an initial learning rate of 0.0001, applying weight decay. Image augmentation is limited to random flipping and cropping. We use a fixed threshold of 0.5 to extract binary masks from the feature map, consistent with previous methods. Hyperparameters α and β are set to 1.0 and 0.1, respectively, for weakly supervised training, and 1.0 and 0.3 for unsupervised training. For the clustering algorithms, we used K-means [34].

Method	CASIAv1		Columbia		Coverage		NIST16		IMD2020	
	IOU	P-F1	IOU	P-F1	IOU	P-F1	IOU	P-F1	IOU	P-F1
NOI [38]	0.075	0.132	0.152	0.236	0.122	0.210	0.048	0.074	**0.091**	0.126
CFAl [17]	0.081	0.134	0.175	0.275	0.103	0.185	0.076	0.105	0.068	0.103
MCA [1]	0.049	0.089	0.085	0.148	0.078	0.136	0.049	0.074	0.044	0.079
NoisePrint [9]	0.074	0.130	0.085	0.320	0.098	0.176	0.062	0.106	0.054	0.104
IVC [8]	0.056	0.101	0.085	0.164	0.070	0.127	0.038	0.068	0.048	0.086
Ours	**0.097**	**0.166**	**0.216**	**0.344**	**0.131**	**0.217**	**0.080**	**0.129**	0.079	**0.136**

Table 1: Evaluation results of unsupervised methods for **Standard Manipulation task**.

Method	CASIAv1				Columbia				Coverage				NIST16		IMD2020	
	IOU	P-F1	I-F1	C-F1	IOU	P-F1	I-F1	C-F1	IOU	P-F1	I-F1	C-F1	IOU	P-F1	IOU	P-F1
FCN [46]	0.078	0.122	0.561	0.200	0.062	0.098	0.524	0.165	0.072	0.122	0.424	0.190	0.032	0.052	0.052	0.086
WSCL [65]	0.100	0.163	0.679	0.263	0.220	0.321	**0.720**	0.444	0.102	0.171	0.571	0.263	0.047	0.078	0.093	0.152
Ours	**0.124**	**0.199**	**0.703**	**0.310**	**0.248**	**0.365**	0.695	**0.479**	**0.140**	**0.221**	**0.667**	**0.332**	**0.079**	**0.131**	**0.124**	**0.204**

Table 2: Evaluation results of weakly supervised approaches for **Standard Manipulation task**.

4.1 Comparison with SoTA Methods

For a fair comparison with SoTA methods, we selected approaches for which the source code is publicly available. Among the unsupervised methods applied for comparison are NOI [38], CFAl [17], MCA [1], NoisePrint [9], and IVC [8], while the weakly supervised methods include FCN [46] and WSCL [65].

Additionally, we conducted experiments using two novel manipulation datasets and compared our approach with fully supervised methods, including RRU-Net [2], Mantra-Net [60], SPAN [25], PSCC-Net [33], Trufor [20], CAT-Net [29], Hifi-Net [21], CR-CNN [62], ObjectFormer [54], and MVSS-Net [5].

Comparison with SoTA unsupervised methods: Due to the assumption of unsupervised methods that all images contain manipulated parts, they will classify all test images as tampered. Thus, they are not suitable for image-level evaluation. We conduct pixel-level experiments to compare their abilities to localize the manipulated region, as shown in Table 1. We can observe that our proposed method in the unsupervised setting achieves the best detection performance compared to other unsupervised methods across five widely used standard manipulation benchmarks.

Comparison with SoTA weakly supervised methods: Table 2 shows experimental results comparing weakly supervised SoTA approaches. Except for the F1 (I-F1) score at the image level in the Columbia [24] dataset, our method performs better than SoTA methods in all other metrics. Regarding the relatively lower I-F1 score in Columbia compared to WSCL [65], we believe the reason is that Columbia does not have post-processing, so our method may not be very sensitive to manipulation. However, despite this issue, our method achieves the best localization performance in the Columbia dataset.

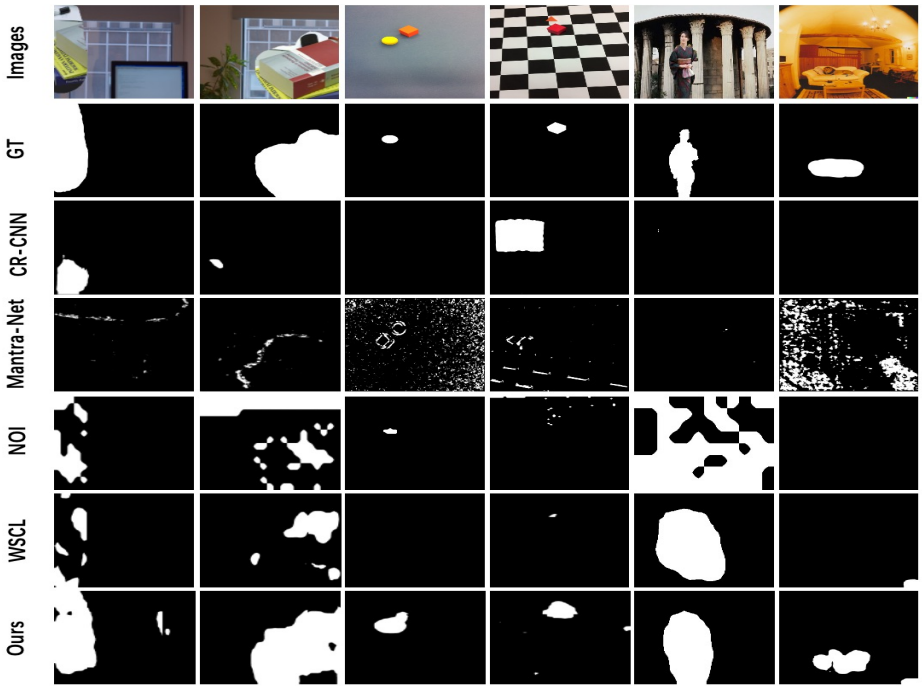

Fig. 4: Visualization results using different methods. The images are displayed in the following order from top to bottom: tampered images, ground truth masks, prediction results from CR-CNN [62], Mantra-Net [60], NOI [38], WSCL [65], and our method.

Comparison using novel manipulation dataset: In order to show the generalization ability of our method. We conduct evaluations using fully supervised and weakly supervised methods on two novel manipulation detection datasets in Table 3. We can see that the fully supervised methods cannot adapt to the novel manipulation types, resulting in low detection performance even if they utilize a very large synthesis training dataset with both image-level and pixel-level labels. In contrast, our method achieves competitive performance while using extremely few training data with only image-level labels.

Visualization Results: We present some visualization results compared to SoTA methods in Figure 4. Our method can better localize the tampered region, even without the use of pixel-level labels. However, due to the lack of pixel-level labels, our model cannot accurately detect tampered edges. These results of our method are generated from the weakly supervised model.

4.2 Ablation Study

We conducted several ablation studies to evaluate the effectiveness of each proposed component. For these studies, we utilized the CASIAv1 [12] and NIST16 [19] datasets.

Type	Method	Training Data Size	IEdit		MagicBrush			
			ACC	I-F1	IOU	P-F1	I-F1	C-F1
Full Supervision	RRU-Net [2]	4.2K	0.482	0.651	0.093	0.153	0.667	0.249
	Mantra-Net [60]	64K	0.499	**0.665**	0.058	0.105	0.667	0.181
	SPAN [25]	96K	0.528	0.210	0.011	0.002	0.585	0.004
	PSCC-Net [33]	100K	0.524	0.206	0.132	0.210	0.710	0.324
	Trufor [20]	858K	0.505	**0.665**	**0.216**	**0.304**	0.670	**0.418**
	CAT-Net [29]	858K	0.488	0.567	0.025	0.033	**0.766**	0.063
	Hifi-Net [21]	1,710K	0.531	0.460	0.089	0.151	0.677	0.247
	CR-CNN [62]	12.5K	0.531	0.530	0.025	0.042	0.593	0.078
	ObjectFormer [54]	12.5K	0.497	0.427	0.029	0.047	0.430	0.085
	MVSS-Net [5]	12.5K	0.526	0.487	0.045	0.072	0.675	0.130
Weak	FCN [46]	12.5K	0.481	0.220	0.020	0.035	0.360	0.064
	WSCL [65]	12.5K	0.511	0.475	0.075	0.122	0.572	0.201
	Ours	12.5K	**0.535**	0.664	0.165	0.264	0.690	0.382

Table 3: Evaluation results on **Novel Manipulation task** for both fully supervised and weakly supervised methods. For the methods that are trained on a dataset with a size of 12.5K, they all utilize CASIAv2 [12] as the training set. For the methods not utilizing CASIAv2, except for RRU-Net, they utilize their synthesis datasets. The best and second-best performances are highlighted using bold and underline, respectively.

Method	CASIAv1			NIST16
	P-F1	I-F1	C-F1	P-F1
Baseline [46]	0.122	0.561	0.200	0.052
Baseline+NRR	0.159	0.572	0.249	0.070
Baseline+NRR+SCL	0.166	0.681	0.267	0.119
Baseline+NRR+SCL+AGAP	**0.199**	**0.703**	**0.310**	**0.131**

Table 4: Ablation study of proposed components using CASIAv1 and NIST16 on Weakly Supervised setting. The baseline in this table is FCN [46].

Method	CASIAv1	NIST16
	P-F1	P-F1
w/o PLS	0.132	0.082
w/ PLS	**0.166**	**0.129**

Table 5: Ablation on pseudo-label selection.

Effectiveness of proposed components: We introduced three novel components: pre-processing stage using Neural Representation Reconstruction (NRR), Selective Pixel-wise Contrastive Learning (SCL), and Adaptive Global-Average Pooling (AGAP) for both un-/weakly supervised IMD. The ablation study conducted in weak mode is shown in Table 4. It is evident that with the progressive integration of our proposed modules, the model's overall ability to detect tampering consistently improves.

Pseudo-Label Selection (PLS): In our unsupervised method, we introduce PLS, which exclusively leverages high-confidence pseudo-labels from two sources to supervise shallow predictions in the self-distillation training process. The impact of PLS is examined in Table 5. In experiments without PLS, we use image-level predictions in the main branch as pseudo-labels to guide shallow

Method	CASIAv1			NIST16
	P-F1	I-F1	C-F1	P-F1
Global-Max Pooling	0.033	0.570	0.062	0.058
Global-Average Pooling	0.158	0.256	0.195	0.081
Generalized Mean Pooling [50]	0.067	0.626	0.121	0.072
Global Smooth Pooling [55]	0.076	0.627	0.136	0.080
Adaptive Global-Average Pooling (Ours)	**0.199**	**0.703**	**0.310**	**0.131**

Table 6: Comparisons using different pooling methods.

predictions. The proposed PLS proves to be effective in enhancing unsupervised performance.

Adaptive Global-Average Pooling: To demonstrate the superiority of the proposed AGAP, we conduct an ablation study in weakly supervised setting using different pooling methods, including Global-Max Pooling (GMP), Global-Average Pooling (GAP), Generalized Mean Pooling (GeM) [50], and Global Smooth Pooling (GsM) [55]. The results are shown in Table 6. Similarly, the proposed AGAP achieves the best performance, highlighting its superiority.

5 Conclusion

We present a novel framework that integrates unsupervised and weakly-supervised approaches for Image Manipulation Detection (IMD). Our approach features a groundbreaking pre-processing step utilizing a controllable fitting function derived from Implicit Neural Representation, providing a prior for manipulation regions. Additionally, we propose a selective pixel-level contrastive learning technique that prioritizes regions with high confidence, mitigating uncertainty stemming from the absence of pixel-level labels. For image-level prediction, we introduce adaptive global average pooling to thoroughly explore manipulation regions for detection and robust training. In unsupervised mode, we implement pseudo-label selection, choosing high-confidence predictions from deeper layers as pseudo-labels to supervise predictions in shallower layers via a self-distillation training method. Extensive experiments validate our method's effectiveness, demonstrating superior performance compared to existing unsupervised and weakly supervised methods. Notably, our approach competes effectively against fully supervised methods in detecting novel manipulations, showcasing its robustness in real-world scenarios. **Limitation** of this work includes inaccurate localization of tampered region edges, resulting in larger prediction masks than the ground truth. **Future work** involves developing more powerful models and effective pre-filters to enhance pixel-level detection performance.

Acknowledgements: This work is supported by the DARPA Semantic Forensics (SemaFor) Program under contract HR001120C0123 and NSF CCSS-2348046.

References

1. Bammey, Q., Gioi, R.G.v., Morel, J.M.: An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14194–14204 (2020) 4, 11
2. Bi, X., Wei, Y., Xiao, B., Li, W.: Rru-net: The ringed residual u-net for image splicing forgery detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019) 11, 13
3. Bondi, L., Lameri, S., Güera, D., Bestagini, P., Delp, E.J., Tubaro, S.: Tampering detection and localization through clustering of camera-based cnn features. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1855–1864. IEEE (2017) 8
4. Chen, K., Hong, L., Xu, H., Li, Z., Yeung, D.Y.: Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7546–7554 (2021) 5
5. Chen, X., Dong, C., Ji, J., Cao, J., Li, X.: Image manipulation detection by multi-view multiscale supervision. In: IEEE/CVF International Conference on Computer Vision. pp. 14185–14193 (2021) 4, 11, 13
6. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8628–8638 (2021) 5
7. Chen, Z., Chen, Y., Liu, J., Xu, X., Goel, V., Wang, Z., Shi, H., Wang, X.: Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2047–2057 (2022) 5
8. Choi, C.H., Choi, J.H., Lee, H.K.: Cfa pattern identification of digital cameras using intermediate value counting. In: Proceedings of the thirteenth ACM multimedia workshop on Multimedia and security. pp. 21–26 (2011) 4, 11
9. Cozzolino, D., Verdoliva, L.: Noiseprint: A cnn based camera model fingerprint. IEEE Transactions on Information Forensics and Security **15**, 144–159 (2019) 4, 11
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 1
11. Dong, J., Wang, W., Tan, T.: CASIA image tampering detection evaluation database. http://forensics.idealtest.org (2010) 2, 10
12. Dong, J., Wang, W., Tan, T.: Casia image tampering detection evaluation database. In: 2013 IEEE China summit and international conference on signal and information processing. pp. 422–426. IEEE (2013) 3, 10, 12, 13
13. Dupont, E., Goliński, A., Alizadeh, M., Teh, Y.W., Doucet, A.: Coin: Compression with implicit neural representations. arXiv preprint arXiv:2103.03123 (2021) 5
14. Ergen, T., Kozat, S.S.: Unsupervised anomaly detection with lstm neural networks. IEEE transactions on neural networks and learning systems **31**(8), 3127–3141 (2019) 2
15. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd. vol. 96, pp. 226–231 (1996) 9
16. Feng, Y., Feng, Y., You, H., Zhao, X., Gao, Y.: Meshnet: Mesh neural network for 3d shape representation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8279–8286 (2019) 5

17. Ferrara, P., Bianchi, T., De Rosa, A., Piva, A.: Image forgery localization via fine-grained analysis of cfa artifacts. IEEE Transactions on Information Forensics and Security **7**(5), 1566–1577 (2012) 4, 11
18. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Transactions on information Forensics and Security **7**(3), 868–882 (2012) 5
19. Guan, H., Kozak, M., Robertson, E., Lee, Y., Yates, A.N., Delgado, A., Zhou, D., Kheyrkhah, T., Smith, J., Fiscus, J.: Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In: IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. pp. 63–72. IEEE (2019) 10, 12
20. Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., Verdoliva, L.: 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. pp. 20606–20615 (2023) 4, 11, 13
21. Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., Liu, X.: Hierarchical fine-grained image forgery detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3155–3165 (2023) 4, 11, 13
22. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 4, 5, 8
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 7, 10
24. Hsu, Y.F., Chang, S.F.: Detecting image splicing using geometry invariants and camera characteristics consistency. In: 2006 IEEE International Conference on Multimedia and Expo. pp. 549–552. IEEE (2006) 2, 10, 11
25. Hu, X., Zhang, Z., Jiang, Z., Chaudhuri, S., Yang, Z., Nevatia, R.: Span: Spatial pyramid attention network for image manipulation localization. In: European Conference on Computer Vision (ECCV),. pp. 312–328. Springer (2020) 4, 11, 13
26. Ji, K., Chen, F., Guo, X., Xu, Y., Wang, J., Chen, J.: Uncertainty-guided learning for improving image manipulation detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22456–22465 (2023) 4, 5
27. Koch, G., Zemel, R., Salakhutdinov, R., et al.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop. vol. 2. Lille (2015) 5
28. Kwan, H.M., Gao, G., Zhang, F., Gower, A., Bull, D.: Hinerv: Video compression with hierarchical encoding-based neural representation. Advances in Neural Information Processing Systems **36** (2024) 5
29. Kwon, M.J., Nam, S.H., Yu, I.J., Lee, H.K., Kim, C.: Learning jpeg compression artifacts for image manipulation detection and localization. International Journal of Computer Vision pp. 1875–1895 (2022) 4, 11, 13
30. Li, J., Chen, Y., Xing, Y.: Memory mechanism for unsupervised anomaly detection. In: The 39th Conference on Uncertainty in Artificial Intelligence (2023) 2
31. Li, S., Xia, X., Ge, S., Liu, T.: Selective-supervised contrastive learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 316–325 (2022) 9
32. Liu, D., Yu, J.: Otsu method and k-means. In: 2009 Ninth International conference on hybrid intelligent systems. vol. 1, pp. 344–349. IEEE (2009) 8
33. Liu, X., Liu, Y., Chen, J., Liu, X.: 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) 4, 11, 13

34. Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982) 10
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 10
36. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3623–3632 (2019) 5
37. Lyu, S., Pan, X., Zhang, X.: Exposing region splicing forgeries with blind local noise estimation. International journal of computer vision **110**, 202–221 (2014) 8
38. Mahdian, B., Saic, S.: Using noise inconsistencies for blind image forensics. Image and vision computing **27**(10), 1497–1503 (2009) 4, 11, 12
39. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021) 7
40. Molaei, A., Aminimehr, A., Tavakoli, A., Kazerouni, A., Azad, B., Azad, R., Merhof, D.: Implicit neural representation in medical imaging: A comparative survey. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2381–2391 (2023) 5
41. Niu, Y., Tondi, B., Zhao, Y., Ni, R., Barni, M.: Image splicing detection, localization and attribution via jpeg primary quantization matrix estimation and clustering. IEEE Transactions on Information Forensics and Security **16**, 5397–5412 (2021) 8
42. Novozamsky, A., Mahdian, B., Saic, S.: Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In: IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. pp. 71–80 (2020) 10
43. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979) 9
44. Pan, X., Zhang, X., Lyu, S.: Exposing image forgery with blind noise estimation. In: Proceedings of the thirteenth ACM multimedia workshop on Multimedia and security. pp. 15–20 (2011) 8
45. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019) 10
46. Pathak, D., Shelhamer, E., Long, J., Darrell, T.: Fully convolutional multi-class multiple instance learning. arXiv preprint arXiv:1412.7144 (2014) 11, 13
47. Pyatykh, S., Hesser, J., Zheng, L.: Image noise level estimation by principal component analysis. IEEE transactions on image processing **22**(2), 687–699 (2012) 8
48. Qian, Y., Hong, X., Guo, Z., Arandjelović, O., Donovan, C.R.: Semi-supervised crowd counting with contextual modeling: Facilitating holistic understanding of crowd scenes. IEEE Transactions on Circuits and Systems for Video Technology (2024) 2
49. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: Learning text-to-image generation by redescription. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1505–1514 (2019) 1
50. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018) 14

51. Shi, J., Xu, N., Bui, T., Dernoncourt, F., Wen, Z., Xu, C.: A benchmark and baseline for language-driven image editing. In: Proceedings of the Asian Conference on Computer Vision (2020) 10
52. Smucny, J., Shi, G., Lesh, T.A., Carter, C.S., Davidson, I.: Data augmentation with mixup: Enhancing performance of a functional neuroimaging-based prognostic deep learning classifier in recent onset psychosis. NeuroImage: Clinical **36**, 103214 (2022) 2
53. Tao, C., Zhu, X., Su, W., Huang, G., Li, B., Zhou, J., Qiao, Y., Wang, X., Dai, J.: Siamese image modeling for self-supervised vision representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2132–2141 (2023) 5
54. Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S.N., Jiang, Y.G.: Objectformer for image manipulation detection and localization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2364–2373 (2022) 4, 11, 13
55. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 136–145 (2017) 14
56. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2314–2320 (2016) 2
57. Wen, B., Zhu, Y., Subramanian, R., Ng, T.T., Shen, X., Winkler, S.: Coverage – a novel database for copy-move forgery detection. In: IEEE International Conference on Image processing (ICIP) (2016) 2, 10
58. Wu, H., Chen, Y., Zhou, J.: Rethinking image forgery detection via contrastive learning and unsupervised clustering. arXiv preprint arXiv:2308.09307 (2023) 4, 8
59. Wu, H., Zhou, J., Tian, J., Liu, J.: Robust image forgery detection over online social network shared images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13440–13449 (2022) 4
60. Wu, Y., AbdAlmageed, W., Natarajan, P.: Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9543–9552 (2019) 4, 11, 12, 13
61. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1316–1324 (2018) 1
62. Yang, C., Li, H., Lin, F., Jiang, B., Zhao, H.: Constrained r-cnn: A general image manipulation detection model. In: IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2020) 4, 5, 11, 12, 13
63. Yang, S., Ding, M., Wu, Y., Li, Z., Zhang, J.: Implicit neural representation for cooperative low-light image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12918–12927 (2023) 3, 5, 7
64. Yoon, J., Yu, S., Bansal, M.: Raccoon: Remove, add, and change video content with auto-generated narratives. arXiv preprint arXiv:2405.18406 (2024) 1
65. Zhai, Y., Luan, T., Doermann, D., Yuan, J.: Towards generic image manipulation detection with weakly-supervised self-consistency learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22390–22400 (2023) 4, 6, 11, 12, 13

66. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. arXiv preprint arXiv:2301.11445 (2023) 5
67. Zhang, H., Wang, R., Zhang, J., Li, C., Yang, G., Spincemaille, P., Nguyen, T., Wang, Y.: Nerd: Neural representation of distribution for medical image segmentation. arXiv preprint arXiv:2103.04020 (2021) 5
68. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36** (2024) 10
69. Zhang, L., Bao, C., Ma, K.: Self-distillation: Towards efficient and compact neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 4388–4403 (2021) 4, 7, 9
70. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 1
71. Zhang, W., Pang, J., Chen, K., Loy, C.C.: Dense siamese network for dense unsupervised learning. In: European Conference on Computer Vision. pp. 464–480. Springer (2022) 5
72. Zhang, Z., Bui, T.D.: Attention-based selection strategy for weakly supervised object localization. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 10305–10311. IEEE (2021) 2
73. Zhang, Z., Chang, M.C.: Two-stage dual augmentation with clip for improved text-to-sketch synthesis. In: 2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 1–6. IEEE (2023) 1
74. Zhang, Z., Chang, M.C., Bui, T.D.: Improving class activation map for weakly supervised object localization. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2624–2628. IEEE (2022) 2
75. Zhang, Z., Li, M., Chang, M.C.: A new benchmark and model for challenging image manipulation detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7405–7413 (2024) 4