Towards Modern Image Manipulation Localization: A Large-Scale Dataset and Novel Methods

Chenfan Qu[1], Yiwu Zhong[2,*], Chongyu Liu[1], Guitao Xu[1], Dezhi Peng[1], Fengjun Guo[3],
Lianwen Jin[1,4,*]
[1]South China University of Technology, [2]University of Wisconsin,
[3]INTSIG Information Co., Ltd
[4]INTSIG-SCUT Joint Lab on Document Analysis and Recognition
202221012612@mail.scut.edu.cn, yzhong52@wisc.edu, eelwjin@scut.edu.cn

Abstract

In recent years, image manipulation localization has attracted increasing attention due to its pivotal role in guaranteeing social media security. However, how to accurately identify the forged regions remains an open challenge. One of the main bottlenecks lies in the severe scarcity of high-quality data, due to its costly creation process. To address this limitation, we propose a novel paradigm, termed as CAAA, to automatically and precisely annotate the numerous manually forged images from the web at the pixel level. We further propose a novel metric QES to facilitate the automatic filtering of unreliable annotations. With CAAA and QES, we construct a large-scale, diverse, and high-quality dataset comprising 123,150 manually forged images with mask annotations. Besides, we develop a new model APSC-Net for accurate image manipulation localization. According to extensive experiments, our dataset significantly improves the performance of various models on the widely-used benchmarks and such improvements are attributed to our proposed effective methods. The dataset and code are publicly available at https://github.com/qcf-568/MIML.

1. Introduction

The rapid development of modern image editing techniques, such as PhotoShop and GIMP, has greatly enriched people's visual world. However, the abuse of manipulated images may lead to fraud and the spread of rumors, posing significant risks to social media security [6, 13]. Consequently, Image Manipulation Localization (IML) has emerged as an important research topic in recent years [11]. It is crucial to develop effective techniques for IML.

The key challenge for IML is the scarcity of manually forged images [35]. Tons of training data are essential for models to prevent overfitting. Unfortunately, the process of elaborately manipulating images and annotating the forged regions at pixel-level is extremely exhaustive and time-consuming. Although synthetic data is utilized to overcome the data scarcity [12, 17], it often exhibits a significant domain gap between real-world manipulations, leading to

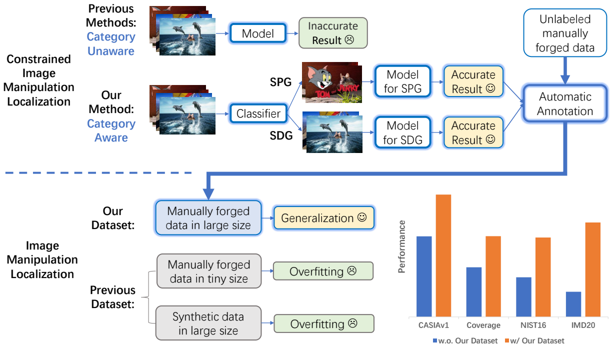
Constrained Image Manipulation Localization
Previous Methods: Category Unaware
Model
Inaccurate Result ☹
Our Method: Category Aware
Classifier
SPG
SDG
Model for SPG
Model for SDG
Accurate Result ☺
Accurate Result ☺
Unlabeled manually forged data
Automatic Annotation
Image Manipulation Localization
Our Dataset:
Manually forged data in large size
Generalization ☺
Previous Dataset:
Manually forged data in tiny size
Overfitting ☹
Synthetic data in large size
Overfitting ☹
Performance
CASIAv1
Coverage
NIST16
IMD20
w.o. Our Dataset
w/ Our Dataset

Figure 1. We propose a novel paradigm for constrained image manipulation localization, which treats images in SPG and SDG separately. We also propose to employ it for auto-annotation and construct a large-scale, high-quality dataset that noticeably enhances the generalization of image manipulation localization models.

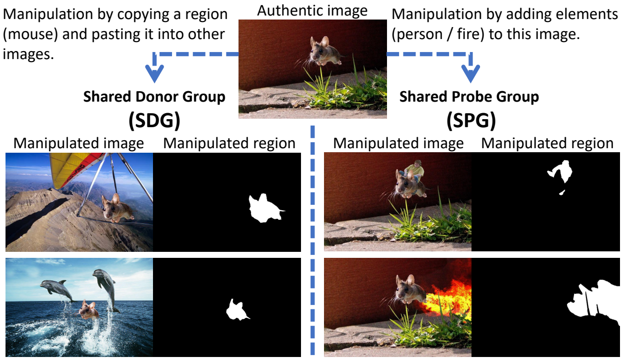
Manipulation by copying a region (mouse) and pasting it into other images.
Authentic image
Manipulation by adding elements (person / fire) to this image.
Shared Donor Group (SDG)
Shared Probe Group (SPG)
Manipulated image
Manipulated region
Manipulated image
Manipulated region

Figure 2. The definition of 'SPG' and 'SDG'. This definition applies to an image pair consisting of a real and a forged image.

[*]: Corresponding author

poor generalization of models trained with it [35]. Recently, various methods have been proposed and successfully improved generalization, such as noise domain features [6, 36] and semantic suppression [27]. However, they are far from fully addressing the issue caused by the severe shortage of manually forged data and still suffer a lot from overfitting.

Considering that there are numerous manually forged images publicly available on the web, along with their corresponding authentic images. We propose a novel idea that employs well-trained constrained image manipulation localization models to automatically obtain the mask annotations for these unlabelled forged images, so that the data scarcity problem for image manipulation localization can be greatly alleviated, as shown in Fig. 1. Since constrained image manipulation localization methods localize forged regions with the help of the corresponding authentic images, the complexity of the task can be considerably reduced.

However, despite progress has been made in less challenging data, previous constrained image manipulation localization methods are inadequate as qualified automatic annotators for complex modern images due to three serious handicaps. First, they mostly employ a single correlation-based model to process all input data [19, 28], which we argue is a suboptimal paradigm. Generally, based on whether the common parts among the manipulated images are forged regions or authentic regions, the pairs of forged images and their original ones can be divided into two groups, Shared Donor Group (SDG) and Shared Probe Group (SPG), as shown in Fig. 2. Although the previous correlation-based methods are reasonable for SDG, they are not suitable enough for SPG, as the actual common parts for data in SPG are mostly background, while those in SDG are mostly foreground. The shared background in SPG have much larger area and much fewer distinctive features compared to the shared foreground in SDG. Simultaneously training correlation-based models on SDG and SPG data will lead to confusion and weaken their generalization ability. Second, the difference maps derived by subtracting the forged images from their authentic ones can always highlight the forged regions, but such an vital clue is completely ignored by previous constrained image manipulation localization methods. Third, previous works don't pay enough attention to the semantic misalignment caused by the substantial re-scaling operation during manipulation, which confuses the models and negatively affects them.

To tackle these problems, we propose a novel paradigm termed as Category-Aware Auto-Annotation (CAAA), which treats image pairs in SDG and SPG separately. The proposed CAAA paradigm consists of three components. Initially, a classifier is employed to determine whether an input image pair belongs to SDG or SPG. This classifier can be trained effectively through self-supervised learning using unlabeled images. Second, a Difference Aware Semantic Segmentation model that utilizes both the image pairs and their difference maps for accurate constrained manipulation localization in SPG. Additionally, a Semantic Aligned Correlation Matching model that improves the performance in SDG through better semantic alignment. Experiments demonstrate that our methods significantly outperform previous constrained image manipulation localization methods on complex scenarios and are adequate for auto-annotation.

Subsequently, we collect a large amount of manually forged images from the Internet and then annotate their forged regions with the proposed CAAA. This approach can considerably alleviate the scarcity of non-synthetic data in image manipulation localization, as shown in Fig. 1. To ensure that all the annotations are reliable enough, we further propose a novel metric, termed as Quality Evaluation Score (QES). The QES can automatically evaluate the quality of the annotations and exclude the bad ones, without needing the ground-truths to calculate. Experiments show that our dataset can significantly improve various image manipulation localization models on the widely-used benchmarks.

Additionally, to make better use of our MIML dataset, we propose a new model, termed as APSC-Net, which outperforms previous methods on various benchmarks.

In summary, our main contributions are as follows:

- We propose a novel idea: facilitating the task of image manipulation localization from the web-scale images and the auto-annotations distilled from a less challenging task, constrained image manipulation localization.
- We propose a novel paradigm for constrained image manipulation localization, termed as CAAA, which treats SPG and SDG separately. For SPG, we propose to employ the image difference denoised with by semantic information. For SDG, we propose to align the semantics with a cross-level feature correlation framework.
- We propose a novel effective metric QES to automatically filter out unreliable mask annotations, during the dataset construction where the ground-truth is not available.
- Based on the above techniques, we construct a large-scale, diverse, high-quality dataset, termed as MIML. It significantly addresses the scarcity of manually forged data for image manipulation localization, thereby considerably improving the generalization ability of the models.

2. Related works

2.1. Image Manipulation Localization

Image manipulation localization aims at localizing the forged regions in images. Due to costly data collection, existing handmade datasets [3, 23] are tiny in size, thereby leading to overfitting in many models [27, 35]. To address overfitting, some studies incorporated handcrafted features in the noise domain. Zhou et al. [36] utilized Steganalysis Rich Model filters to help suppress semantic features unrelated to manipulation operation. Dong et al. [2] further used

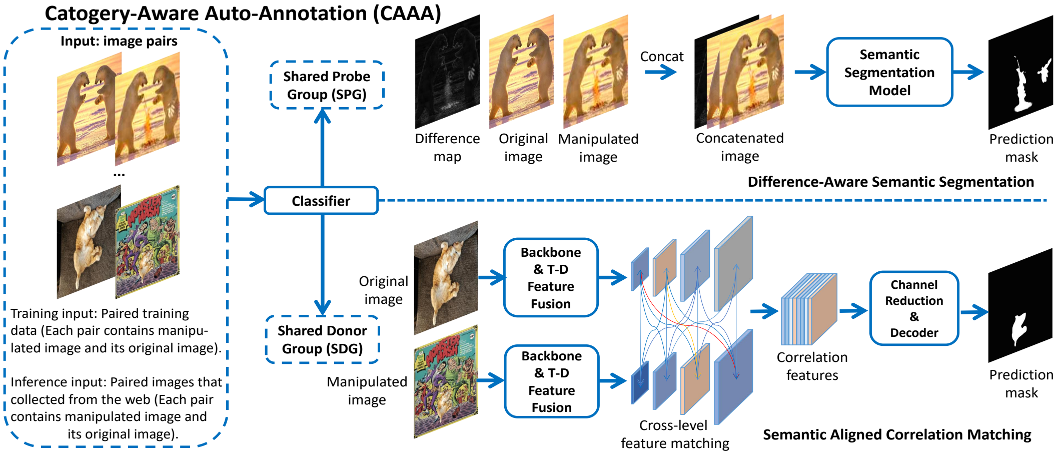
Catogery-Aware Auto-Annotation (CAAA)
Input: image pairs
...
Training input: Paired training data (Each pair contains manipulated image and its original image).
Inference input: Paired images that collected from the web (Each pair contains manipulated image and its original image).
Shared Probe Group (SPG)
Classifier
Shared Donor Group (SDG)
Difference map
Original image
Manipulated image
Concat
Concatenated image
Semantic Segmentation Model
Prediction mask
Difference-Aware Semantic Segmentation
Original image
Manipulated image
Backbone & T-D Feature Fusion
Backbone & T-D Feature Fusion
Cross-level feature matching
Correlation features
Channel Reduction & Decoder
Prediction mask
Semantic Aligned Correlation Matching

Figure 3. The proposed Category-Aware Auto-Annotation (CAAA) paradigm. In this CAAA paradigm, the input image pairs are first fed to a classifier to be classified into SPG or SDG. We then use Difference Aware Semantic Segmentation and Semantic Aligned Correlation Matching to predict manipulated regions in SPG and SDG respectively. The predictions serve as automatic annotations.

Bayar filter for learnable noise domain modeling. Kwon et al. [12] proposed using double JPEG compression artifacts to localize manipulations. Wang et al. [29] employed high-pass filters for anomaly detection. However, the hand-crafted features are mostly noisy and unstable, hindering further improvements in model performance. Recently, pure vision models with semantic suppression techniques have achieved better performance. Sun et al. [27] proposed semantic-agnostic feature learning framework to reduce the bias introduced by semantic features. Zhou et al. [35] employed contrastive learning loss to improve generalization ability. Despite the progress made, these works still suffer a lot from over-fitting due to inadequate high-quality data. To this end, we propose to address data scarcity by constructing a large-scale, diverse, and high-quality dataset, with the Internet images accurately annotated in an automatic way.

2.2. Constrained Image Manipulation Localization

In contrast to image manipulation localization, constrained image manipulation localization (CIML) [32] localizes the forged image regions with the extra help of the given authentic image. Most of the previous works were based on correlation matching, and treated image pairs in SDG and SPG uniformly. Wu et al. [32] proposed DMVN, the first deep correlation model, which computed correlation maps to localize similar objects in images. Liu et al. [19] proposed to remove the pooling layers and adopted atrous convolution for richer spatial information. Liu et al. [18] employed attention-aware mechanism for better performance. Tan et al. [28] proposed performing correlation in both the encoder and decoder to extract better features. These methods achieved significant progress on the datasets that are less challenging (e.g., synthetic COCO [32]). However, their performance are limited in modern images that have high resolution, large variation, and great complexity.

3. Category-Aware Auto-Annotation

For constrained image manipulation localization, previous works didn't consider the discrepancy between SPG and SDG image pairs, and processed them uniformly using a single correlation-based model. We argue that such a paradigm is sub-optimal and the reasons are as follows:

First, the similar regions for SDG images are foreground (*e.g.*, the cats in the SDG branch of Fig. 3). They have specific, similar shapes and unique features. In contrast, the similar regions for SPG images are background. They don't usually have features distinctive enough for accurate correlation matching (*e.g.*, in the images of the SPG branch in Fig. 3, a patch of snow from the background has high similarity to all other patches of snow in the background). Therefore, these regions are likely to cause confusion in the correlation-based models, especially in complex scenarios.

Second, the difference between the paired SPG images is an important cue. Most of the area in an SPG image pair is almost the same and spatially aligned (*e.g.*, the image pair of SPG branch in Fig. 3). Simply subtracting between them and the resulting difference map can always highlight the manipulated regions. However, such information is difficult to be utilized in the previous correlation matching based models and thus not considered in previous CIML works.

Based on these observations, we propose a new paradigm for CIML task, Category-Aware Auto-Annotation. The key idea is to process SPG and SDG images independently, as shown in Fig. 3. First, the input image pairs are categorized into SPG or SDG with a classifier proposed in Section 3.1. For SPG, image pairs are processed by the Difference Aware Semantic Segmentation proposed in Section 3.2. For SDG, image pairs are processed by the Semantic Aligned Correlation Matching proposed in Section 3.3.

More importantly, the models trained with our proposed paradigm are further utilized to perform automatic annota-

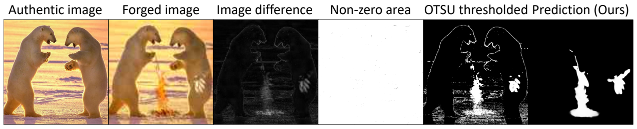

Figure 4. Since manipulated images usually undergo a series of degradation during transmission, the absolute difference between them and their authentic images cannot accurately indicate the forged regions. Our method achieves adequate denoising through the use of semantic information, thereby solving this problem.

tion on numerous manually forged images from the web. In return, the collected data addresses the severe scarcity of non-synthetic data for image manipulation localization.

3.1. Self-Supervised Classification

To achieve the classification of SPG and SDG, we propose to train a classifier via self-supervised learning with unlabelled images. Given an image, we perform random augmentations and manipulations on it, then the forged image and the original image form an SPG pair. To construct an SDG pair, we copy random objects from the original image, resize them, and paste them into another image. With the obtained image pairs, we can effectively train our classifier. The two images in each input pair are concatenated in the channel dimension before being fed into the classifier. The classifier only needs to discern whether the two images in an image pair are almost the same (SPG) or clearly different (SDG), without considering which or where is fake. Therefore, this classification task is quite simple and we can accurately separate the image pairs into the two groups.

3.2. Difference-Aware Semantic Segmentation

Ideally, for an image pair in SPG, the absolute difference between the authentic image and the forged image is actually the forged region. However, manipulated images typically suffer degradation during transmission [31], making it unfeasible to utilize the absolute difference as a precise annotation. As shown in Fig. 4. almost all of the area in the image difference map is non-zero due to transmission degradation. Even the difference map binarized by the OTSU [24] algorithm exhibits highlights on authentic regions, particularly in high-frequency area such as edges. To address this issue, we propose to denoise the difference map by utilizing the semantic information from the images. To achieve this, we propose inputting the channel dimension concatenation of the authentic image, the forged image and their difference map into a semantic segmentation model.

3.3. Semantic Aligned Correlation Matching

Due to the extensive rescaling operations, semantic misalignment becomes a key factor that has an adverse impact on the effectiveness of correlation-based methods. For example, in the SDG branch of Fig. 3. the cat in the original image occupies a large region, whereas in the forged image, the same cat is confined to a much smaller one. The cat features of the original image are mostly in the highest level but those of the forged image are mostly in the lowest level. Hence, the visual features at the same encoding level between the two images have misaligned semantics. However, previous works simply force the models to perform feature matching between the same feature level, which confuses the models and negatively affects their generalization. To this end, we propose to improve the performance of correlation model by achieving better semantic alignment.

Specifically, given a set of feature maps with different resolutions extracted from the backbone model, we first compute global representations from the highest features with average pooling, and then fuse them with the highest features with a convolutional layer. subsequently, we fuse these feature maps in a top-down manner similar to that in FPN [15, 34]. In this way, the low-level features have more semantics and are prepared to match the high-level features. We then calculate correlation features F_{corr} between the features of the input image pair in a cross-level manner as equation (1), which differs from previous methods [18, 19, 32] that compute correlation features solely between feature maps of the same level as equation (2).

$$[Corr(F_{o,i}, F_{m,j})\; for\; i\; in\; (0\text{–}3)\; and\; for\; j\; in\; (0\text{–}3)] \quad (1)$$

$$[Corr(F_{o,i}, F_{m,i})\; for\; i\; in\; (0\text{–}3)\,] \quad (2)$$

In these equations, $Corr$ denotes the correlation function widely-used in previous works [18, 19, 32], $F_{o,i}$ denotes the ith level feature map from the original image and $F_{m,j}$ denotes the jth level feature map from the forged image. Our model is able to adaptively select the optimal matching route, leading to enhanced semantic alignment. F_{corr} is subsequently concatenated, channel reduced and fed into a convolutional decoder for the final prediction.

4. MIML Dataset

In this section, we propose a large-scale, diverse, high-quality dataset, termed as MIML. The key idea is leveraging the constrained image manipulation localization models trained on existing datasets to automatically obtain accurate mask annotations for the manually forged images from the web. To ensure the dataset is of high quality, we also propose a novel metric to filter out the inadequate annotations.

4.1. Dataset Construction

As shown in Fig 6, we construct MIML as follow steps:
Image Collection. We collect image pairs from imgur.com. On this website, the images are manually forged by millions of people and thus have high-quality, diverse forged regions.
Data Clean. We clean the collected data and exclude the images that overlap with the evaluation datasets in Sec.6.
Classification We categorize the cleaned image pairs into SPG or SDG using the classifier proposed in Sec.3.1. The actual classifier is an ensemble of three models [8, 20, 21].

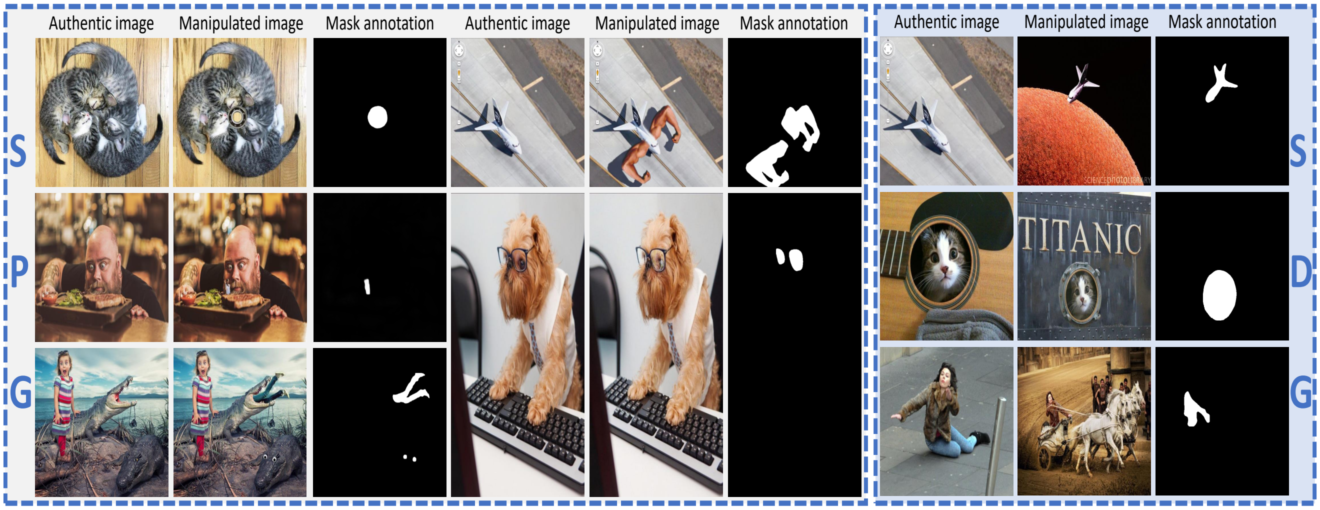

Figure 5. Some of the images and their corresponding binarized mask annotations from the proposed MIML dataset.

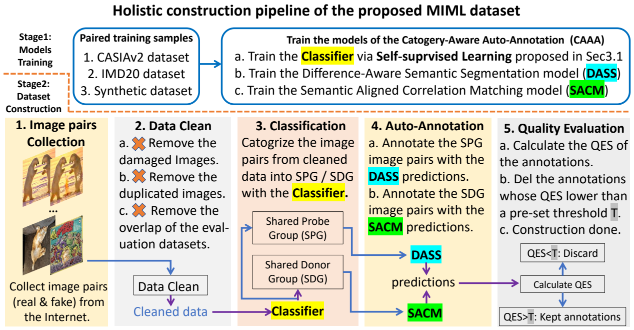
Holistic construction pipeline of the proposed MIML dataset
Stage1: Models Training
Stage2: Dataset Construction
Paired training samples
1. CASIAv2 dataset
2. IMD20 dataset
3. Synthetic dataset
Train the models of the Catogery-Aware Auto-Annotation (CAAA)
a. Train the Classifier via Self-suprvised Learning proposed in Sec3.1
b. Train the Difference-Aware Semantic Segmentation model (DASS)
c. Train the Semantic Aligned Correlation Matching model (SACM)
1. Image pairs Collection
...
Collect image pairs (real & fake) from the Internet.
2. Data Clean
a. ✗ Remove the damaged Images.
b. ✗ Remove the duplicated images.
c. ✗ Remove the overlap of the evaluation datasets.
Data Clean
Cleaned data
3. Classification
Catogrize the image pairs from cleaned data into SPG / SDG with the Classifier.
Shared Probe Group (SPG)
Shared Donor Group (SDG)
Classifier
4. Auto-Annotation
a. Annotate the SPG image pairs with the DASS predictions.
b. Annotate the SDG image pairs with the SACM predictions.
DASS
predictions
SACM
5. Quality Evaluation
a. Calculate the QES of the annotations.
b. Del the annotations whose QES lower than a pre-set threshold T.
c. Construction done.
QES<T: Discard
Calculate QES
QES>T: Kept annotations

Figure 6. The construction pipeline of our MIML dataset.

Auto-Annotation. We utilize the DASS and SACM proposed in Sec.3.2 and Sec.3.3 to automatically obtain mask annotations for the images in SPG and SDG respectively.

Quality Evaluation. After auto-annotation, the annotations of SPG already have high quality, while a few annotations of SDG are still unsatisfactory. To ensure overall quality, we propose a novel metric, Quality Evaluation Score (QES), to further filter out the unreliable annotations. The key idea of QES is that most of the high quality predictions have very high confidence and sharp edges, thus we can evaluate the predictions' quality and exclude the bad ones by examining their confidence and sharpness. Specifically, given a prediction mask with shape (H, W) and normalized probability, we compute the QES as follows: QES=$\frac{\sum_{i,j}^{H,W} p_{i,j} > (1-T_h)}{\sum_{i,j}^{H,W} p_{i,j} > T_l}$, where $\sum_{i,j}^{H,W} p_{i,j} > (1-T_h)$ denotes the area of prediction with a high confidence greater than $(1-T_h)$, and $\sum_{i,j}^{H,W} p_{i,j} > T_l$ denotes the total predicted potentially manipulated area. We set T_h and T_l to $\frac{1}{16}$ and only retain samples with QES>0.5. Experiments show that our QES has strong correlation with the IoU metric and can effectively assist in filtering out the unreliable mask annotations.

Name	Year	Nums	(Height, Width) Range
CASIAv1 [3]	2013	921	(246, 384)-(500, 334)
CASIAv2 [3]	2013	5,123	(160, 240)-(901, 600)
Coverage [30]	2016	100	(190, 334)-(472, 752)
NIST16 [5]	2016	564	(500, 500)-(3744, 5616)
In Wild [10]	2018	201	(650, 650)-(2736, 3648)
IMD20 [23]	2020	2,010	(193, 260)-(4437, 2958)
MIML (Ours)	**2024**	**123,150**	**(45, 120)-(13846, 9200)**

Table 1. A brief summary of previous publicly available handcrafted datasets for IML. Some handcrafted datasets with less than 3k samples and rarely used are omitted. 'Nums' denotes the number of annotated forged samples in the dataset.

4.2. Dataset Highlights

We present a few examples of the proposed dataset in Fig. 5. The main highlights of our dataset are as follows:

- **High quality.** Image manipulation in the proposed dataset is elaborately crafted by humans. Such data can teach models to spot forgery in real-world scenarios, rather than merely overfitting a few simple patterns in synthetic data.
- **Large Scale.** As shown in Table 1, the proposed dataset has a total of 123,150 manually forged images, which is dozens of times more than the previous handmade IML datasets (*e.g.*, ≈60 times more samples than IMD20).
- **Board Diversity.** Our dataset comprises images of various sizes, various styles and various types of manipulation (*e.g.*, copy-move, splicing, removal). They are created by hundreds of thousands of individuals utilizing various software. Such diverse data can considerably enhance the generalization ability of deep IML models.
- **Modern Style.** Our dataset has a large number of modern images that were recently captured and forged, keeping up with modern technology for digital photography. In contrast, the CASIA dataset [3] was proposed more than

a decade ago, where most images have small sizes and are blurred. Therefore, our dataset can better meet the requirements of modern image manipulation localization.

- **Strong Scalability.** There are many increasingly popular image manipulation competitions on the web, which continuously attract millions of people to take in for entertainment (*e.g.* 19 million people in PS-Battles [9, 25]), resulting in numerous new manually forged images. Our dataset construction approach is ready to harness these growing cheap web data. Therefore, our dataset can be easily expanded, demonstrating strong scalability.

5. APSC-Net

In this section, we propose a new model, named APSC-Net, to achieve accurate image manipulation localization. As shown in Fig. 7, it consists of a feature extractor, an Adaptive Perception module and a Self-Calibration module.

5.1. Adaptive Perception Module

During meticulous image forensic analysis, humans often zoom in and out the image repeatedly, selecting an optimal set of observations to assist their final prediction. To mimic the human perception way, we design an Adaptive Perception module to help the model compare between different views and adaptively select the optimal combination for each input image. The key idea is to weighted sum the current and all the higher-level feature maps using adaptive weights calculated from the their global representations.

Specifically, given four feature maps extracted from the backbone model, we first map them to the same number of channels with an $1{\times}1$ conv-layer and obtain four feature maps $F_{i,0}, F_{i,1}, F_{i,2}, F_{i,3}$. We then get global image representations from $F_{i,3}$ with global average pooling, and fuse them with $F_{i,3}$ to $F_{o,3}$ using an $1{\times}1$ conv-layer. Finally, for a in $(2, 1, 0)$ and for b in range $(a + 1, 3)$, we successively calculate $F_{o,a}$ following the equation (3) and (4) below:

$$[w_{a,a}, w_{a,b}] = \sigma(f_a(Cat([Avg(F_{i,a}), Avg(F_{o,b})]))) \quad (3)$$

$$F_{o,a} = Conv(w_{a,a} * F_{i,a} + \sum_{b=a+1}^{3} w_{a,b} * F_{o,b}) \quad (4)$$

where Avg denotes global average pooling, Cat denotes channel dimension concatenation, f_a denotes two linear layers with a ReLU [4] layer, σ denotes the Sigmoid activation function and $Conv$ denotes a $3{\times}3$ conv-layer.

5.2. Self Calibration Module

When performing meticulous localization of manipulated images, humans are inclined to confirm their initial predictions by comparing the features surrounding the predicted forged regions. Additionally, they might amend their local prediction based on their global evaluation of the image's authenticity. To emulate the human perception way, we design a Self Calibration module for better performance. As shown in Fig. 7, the proposed Self Calibration module comprises a Segmentation-based Self Calibration (SSC) and a Classification-based Self Calibration (CSC).

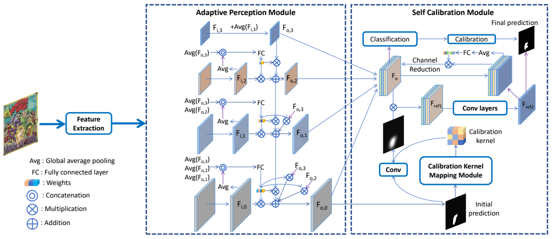

Figure 7. The overall framework of the proposed APSC-Net.

For the SSC, the initial prediction from the end of the Adaptive Perception module is obtained and fed into a tiny Calibration Kernel Mapping Module, which consists of several convolutional layers. Subsequently, we get a calibration kernel and conduct convolutional operation on the initial prediction with it. The resulting values are then normalized using a Min-Max approach. We multiply the normalized result by F_o, the concatenation of $F_{o,0}, F_{o,1}, F_{o,2}, F_{o,3}$, and get F_{ref1}. Next, we refine F_{ref1} with several convolutional layers and get the refined feature F_{ref2}. After that, we concatenate F_{ref2} with F_o, perform channel attention and channel reduction, substitute F_o with the outcome, and repeat the process that utilizes F_o and calibrated prediction to obtain F_{ref2} twice again, for a refined version of F_{ref2}. Then the resulting F_{ref2} is utilized for final prediction. With the SSC, our model can adaptively attend to the optimal region roughly based on its initial mask prediction, thereby achieves higher performance via in-depth analysis.

For the CSC, we begin by feeding the refined features F_{ref2} into a tiny classification head that predicts whether the input image is manipulated or not. If the image is predicted as authentic, the mask prediction is likely to have more false positive (FP), so we increase the binarization threshold to reduce the FP. On the other hand, if the image is predicted as manipulated, we decrease the binarization threshold to reduce the false negative. Given a probability P that the input image is predicted as forged, CSC adjusts the binarization threshold of the prediction mask from 0.5 to $min(max(1 - P, \lambda), 1 - \lambda)$, and λ is set to 0.3.

6. Experiments

6.1. Experiments for CIML

The task of image manipulation automatic annotation can be evaluated as a CIML task. Considering that images in the IMD20 dataset [23] have very similar style to the target images we aim to annotate, we use part of them to evaluate models' performance with IoU and F1-score.

Implementation Details. We categorize the forged images in IMD20 into SPG or SDG and randomly split them into training and testing sets with an approximate 3:1 ratio. The CASIAv2 [3] and about one million images synthesized

Dataset	PSCC-Net [17]						CAT-Netv2 [12]						APSC-Net (Ours)					
	IoU			F1			IoU			F1			IoU			F1		
	w.o.	w/	**gain**	w.o.	w/	**gain**	w.o.	w/	**gain**	w.o.	w/	**gain**	w.o.	w/	**gain**	w.o.	w/	**gain**
CASIAv1	.401	.609	+52%	.430	.649	+51%	.660	.691	+5%	.703	.728	+4%	.799	.810	+1%	.837	.848	+1%
NIST16	.247	.402	+62%	.295	.476	+61%	.239	.353	+48%	.287	.422	+47%	.398	.525	+35%	.436	.590	+35%
Coverage	.197	.395	+100%	.218	.477	+118%	.245	.302	+23%	.286	.389	+36%	.490	.498	+2%	.523	.568	+8%
IMD20	.125	.470	+277%	.156	.541	+247%	.157	.547	+248%	.192	.629	+228%	.339	.679	+101%	.391	.760	+95%

Table 2. IML ablation study for the proposed MIML dataset. 'w.o.' denotes training without the MIML dataset, 'w/' denotes training with the MIML dataset, and 'gain' denotes the ratio of improvement in performance.

QES Threshold	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Number ratio of kept predictions	3.03	1.90	1.66	1.43	1.21	1.00	0.78	0.54	0.29	0.06
Kept predictions' IoU on IMD20SDG	.702	.812	.825	.851	.881	.903	.917	.948	.966	.968

Table 3. Ablation study for the threshold of the proposed Quality Evaluation Score (QES).

Methods	IoU	F1
DMVN* [32]	.276	.430
DMVN [32]	.578	.728
DMAC* [19]	.432	.620
DMAC [19]	.573	.728
Ours(VGG [26])	.781	.860
Ours(VAN [7])	**.834**	**.889**

Methods	IoU	F1
Nonzero	.079	.273
OTSU [24]	.497	.683
w.o. Difference	.718	.827
w.o. Images	.741	.812
Ours(VGG [26])	.781	.860
Ours(VAN [7])	**.834**	**.889**

Table 4. CIML experiments on the IMD20 SPG. Left: Comparison study for our Difference-Aware Semantic segmentation. Right: Ablation study for it. 'DMVN*' denotes DMVN trained with both SDG and SPG data, akin to 'DMAC*'. 'Nonzero' denotes using the non-zero region of the difference between a pair of images, 'OTSU' denotes the difference binarized with OTSU. 'w.o. Difference' denotes that the input of the semantic segmentation model contains only the image pairs, 'w.o. Images' denotes that only using the difference map of image pairs as input. 'Ours(VGG)' denotes our model with the same VGG backbone as DMAC. 'Ours(VAN)' denotes our model with the VAN backbone.

Methods	IoU	F1
DMVN* [32]	.276	.434
DMVN [32]	.317	.495
DMAC* [19]	.410	.559
DMAC [19]	.518	.660
Ours	**.702**	**.798**

Set	TDF	CLM	QES	IoU	F1
(1)				.590	.731
(2)	✓			.608	.741
(3)	✓	✓		.702	.798
(4)	✓	✓	✓	**.903**	**.950**

Table 5. CIML experiments on the IMD20 SDG. Left: Comparison study for our Semantic Aligned Correlation Matching. Right: Ablation study for it. 'DMVN*' denotes DMVN trained with both SDG and SPG data, similar to 'DMAC*'. 'TDF' denotes Top-Down Fusion, 'CLM' denotes Cross-Level Matching, 'QES' denotes filtering predictions with our Quality Evaluation Score and the evaluation includes only the predictions with QES > 0.5.

Setting	Training-Set+		CASIAv1		NIST16		IMDP2	
	IMDP1	MIML	IoU	F1	IoU	F1	IoU	F1
(1)			.799	.837	.398	.436	.351	.402
(2)	✓		.790	.825	.431	.479	.590	.667
(3)		✓	**.810**	**.848**	**.525**	**.590**	**.703**	**.788**

Table 6. IML ablation study for our MIML, 'Training Set+' denotes the training set except for CASIAv2 and synthetic data.

Method	CASIAv1		NIST16		Coverage		Columbia	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1
ManTraNet [33]	.086	.130	.040	.062	.181	.271	.274	.377
RRU-Net [1]	.330	.380	.080	.129	.165	.260	.351	.476
MVSS-Net [2]	.403	.455	.243	.294	.389	.454	.578	.668
PSCC-Net [17]	.410	.463	.067	.110	.340	.446	.469	.603
CAT-Netv2 [12]	.684	.738	.238	.302	.238	.292	.760	.804
IF-OSN [31]	.465	.509	.247	.326	.181	.268	.610	.710
EVP [16]	.438	.502	.188	.239	.078	.114	.200	.263
TruFor [6]	.630	.692	.279	.348	.446	.522	.748	.812
Ours(w.o. MIML)	.799	.837	.398	.436	.490	.523	.956	.966
Ours(w/ MIML)	**.810**	**.848**	**.525**	**.590**	**.498**	**.568**	**.962**	**.973**

Table 7. IML comparison study for the proposed APSC-Net.

with the COCO dataset [14] are also used for training. Input image is resized to 512x512, and consistent training configuration is applied across all methods for fair comparisons.

Ablation Study. For SPG, the image difference between forged images and their authentic ones can roughly indicate the forged regions, the images themselves can provide semantic information to help the model denoise the difference maps. We conduct ablation study for the proposed Difference Aware Semantic Segmentation on the testing set of IMD20 SPG, as shown in the right part of Table 4, both of these methods can enhance the model's performance. For

SDG, semantic alignment can reduce the confusion during training and help our model achieve better generalization. We conduct ablation study for the proposed Semantic Aligned Correlation Matching on the testing set of IMD20 SDG, the results are shown in the right part of Table 5. Obviously that both the proposed components contribute towards a higher performance of the model. Moreover, the proposed Quality Evaluation Score (QES) allows for the automatic filtering of the most satisfactory predictions. As there are some errors in IMD20's ground truth, our methods are sufficient for obtaining accurate automatic annotations.

Ablation Study for QES. The goal of the proposed QES **metric** is to automatically filter out bad predictions during dataset creation, where the ground truth is not available. As shown in Table 3, a higher QES threshold **always** leads to higher accuracy. This is because predictions with a larger ratio of high confidence and sharper edges are mostly closer to the actual ground truth, and the sharpness and confidence can be well evaluated by our QES. Consequently, our QES demonstrates a strong correlation with the IoU metric.

Comparison Study. We re-train the DMVN [32] and DMAC [19] with their public codes on the same data as ours, the results are shown in the left part of Table 4 and Table 5. Obviously our methods significantly outperform these previous methods. It's notable that DMVN and DMAC trained with both SPG and SDG data perform worse on both tasks than those trained with only SPG or SDG data.

6.2. Experiments for IML

Implementation Details. We adopt ConvNeXt-Base [21] as the feature extractor and train the model for 160k iterations with a batch size of 20, The input size is set to 512x512 during training following the previous works [6, 12]. We use Cross-Entropy loss and AdamW optimizer [22] with a learning rate linearly decaying from 1e-4 to 1e-6. The CASIAv2 [3] and the synthetic dataset as in CAT-Net [12] are used for training following the previous works [6, 12].

Ablation Study for MIML dataset. Except for our APSC-Net, we re-train PSCC-Net [17] and CAT-Net [12] with their public codes with and without the proposed MIML dataset respectively. When training with the MIML dataset, we adopt an approximate 1:1 sampling ratio for the original synthetic data and MIML, the total training volume is fixed in all experiments for fair comparisons. As shown in Table 2, **MIML can significantly improve the performance of all these models on commonly used real-world benchmarks without any additional burden during training or testing**. This is because MIML can greatly alleviate the severe shortage of manually forged data for deep IML models. To further confirm the effectiveness of our MIML dataset, we randomly divide the IMD20 dataset into IMDP1 with 1012 samples and IMDP2 with 988 samples, substitute the MIML dataset with augmented IMDP1 of a similar scale and train APSC-Net with them. As shown in Table 6, despite the inclusion of IMDP1 in the training mitigates the domain gap and improves the performance of the model on NIST16 and IMDP2, it is still significantly worse than the one trained with MIML. Obviously MIML can notably enhance the generalization ability of deep models with its considerable volume of diverse manually forged data.

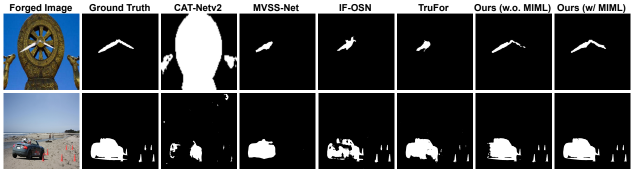

Figure 8. Qualitative results on CASIAv1 and NIST16 datasets.

Comparison Study for APSC-Net. We compare the performance of our APSC-Net with state-of-the-art (SOTA) methods on the widely-used benchmarks. Considering that the previous methods performed different post-processing, which leads to unfairness (*e.g.* EVP [16] used the best threshold calculated from GT to perform binarization), we ignore the post-processing unrelated to their proposed methods and uniformly binarize the predictions with a fixed threshold 0.5, then evaluate the performance with the vanilla IoU and F1-score metrics. The quantitative results are shown in Table 7, our APSC-Net outperforms previous state-of-the-art methods on all these benchmarks. The qualitative results for visual comparison are shown in Fig. 8.

7. Conclusion

In this paper, we propose a novel paradigm for Constrained Image Manipulation Localization (CIML), termed as CAAA, which treats Shared Probe Group and Shared Donor Group image pairs separately. Experiments show that the proposed paradigm considerably outperforms previous CIML methods. With this paradigm, the trained models are used to automatically annotate unlabeled forged images for image manipulation localization. We also propose a novel metric QES to automatically exclude bad predictions. As a result, we harvest a large-scale, diverse, high-quality dataset MIML, with 123,150 manually forged images and pixel-level annotations, which can inspire the potential of deep forensic models by addressing their data scarcity issue. Further, we propose a new effective model APSC-Net for image manipulation localization. We hope our proposed CAAA paradigm, QES metric, MIML dataset and APSC-Net can bring insights to the community and promote the real-world application of image manipulation localization.

Acknowledgement Acknowledgement This paper is supported in part by National Natural Science Foundation of China (Grant No.: 61936003, App. No.: 6244100044), National Key Research and Development Program of China (2022YFC3301703).

References

[1] Xiuli Bi, Yang Wei, Bin Xiao, and Weisheng Li. Rru-net: The ringed residual u-net for image splicing forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 7

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

[3] 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. 2, 5, 6, 8

[4] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011. 6

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

[6] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20606–20615, 2023. 1, 2, 7, 8

[7] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational Visual Media*, 9(4):733–752, 2023. 7

[8] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6185–6194, 2023. 4

[9] Silvan Heller, Luca Rossetto, and Heiko Schuldt. The PS-Battles Dataset – an Image Collection for Image Manipulation Detection. *CoRR*, abs/1804.04866, 2018. 6

[10] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[11] Kaixiang Ji, Feng Chen, Xin Guo, Yadong Xu, Jian Wang, and Jingdong Chen. Uncertainty-guided learning for improving image manipulation detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22456–22465, 2023. 1

[12] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. 1, 3, 7, 8

[13] Dong Li, Jiaying Zhu, Menglu Wang, Jiawei Liu, Xueyang Fu, and Zheng-Jun Zha. Edge-aware regional message passing controller for image forgery localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8222–8232, 2023. 1

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[16] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19434–19445, 2023. 7, 8

[17] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7505–7517, 2022. 1, 7, 8

[18] Yaqi Liu and Xianfeng Zhao. Constrained image splicing detection and localization with attention-aware encoder-decoder and atrous convolution. *IEEE Access*, 8:6729–6741, 2020. 3, 4

[19] Yaqi Liu, Xiaobin Zhu, Xianfeng Zhao, and Yun Cao. Adversarial learning for constrained image splicing detection and localization based on atrous convolution. *IEEE Transactions on Information Forensics and Security*, 14(10):2551–2566, 2019. 2, 3, 4, 7, 8

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[21] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4, 8

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*, pages 10012–10022, 2019. 8

[23] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2020. 2, 5, 6

[24] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. 4, 7

[25] Reddit. Photoshopbattles. https://www.reddit.com/r/photoshopbattles/. 6

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[27] Zhihao Sun, Haoran Jiang, Danding Wang, Xirong Li, and Juan Cao. Safl-net: Semantic-agnostic feature learning network with auxiliary plugins for image manipulation detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22424–22433, 2023. 2, 3

[28] Yuxuan Tan, Yuanman Li, Limin Zeng, Jiaxiong Ye, Xia Li, et al. Multi-scale target-aware framework for constrained image splicing detection and localization. *arXiv preprint arXiv:2308.09357*, 2023. 2, 3

[29] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022. 3

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

[31] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2022. 4, 7

[32] 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 *Proceedings of the 25th ACM international conference on Multimedia*, pages 1480–1502, 2017. 3, 4, 7, 8

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

[34] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 4

[35] Jizhe Zhou, Xiaochen Ma, Xia Du, Ahmed Y Alhammadi, and Wentao Feng. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22346–22356, 2023. 1, 2, 3

[36] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1053–1061, 2018. 2