

Omni-IML: Towards Unified Image Manipulation Localization

Chenfan Qu¹, Yiwu Zhong^{2,*}, Fengjun Guo³, Lianwen Jin^{1,*}

¹South China University of Technology, ²University of Wisconsin, ³INTSIG Information Co., Ltd

202221012612@mail.scut.edu.cn, yzhong52@wisc.edu, eelwjin@scut.edu.cn

Abstract

Image manipulation can lead to misinterpretation of visual content, posing significant risks to information security. Image Manipulation Localization (IML) has thus received increasing attention. However, existing IML methods rely heavily on task-specific designs, making them perform well only on one target image type but are mostly random guessing on other image types, and even joint training on multiple image types causes significant performance degradation. This hinders the deployment for real applications as it notably increases maintenance costs and the misclassification of image types leads to serious error accumulation. To this end, we propose *Omni-IML*, the first generalist model to unify diverse IML tasks. Specifically, *Omni-IML* achieves generalism by adopting the *Modal Gate Encoder* and the *Dynamic Weight Decoder* to adaptively determine the optimal encoding modality and the optimal decoder filters for each sample. We additionally propose an *Anomaly Enhancement* module that enhances the features of tampered regions with box supervision and helps the generalist model to extract common features across different IML tasks. We validate our approach on IML tasks across three major scenarios: natural images, document images, and face images. Without bells and whistles, our *Omni-IML* achieves state-of-the-art performance on all three tasks with a single unified model, providing valuable strategies and insights for real-world application and future research in generalist image forensics. Our code will be publicly available.

1. Introduction

The rapid advancement of image processing software and deep generative models has considerably enriched human capability to create innovative visual content. Users can effortlessly manipulate the visual appearance and create new images that do not exist [27]. Inevitably, such forged images can lead to fraud and the spread of rumors, posing significant risks to politics, economics, and personal privacy [26]. Consequently, Image Manipulation Localization (IML) has become an emerging issue in social media security [30].

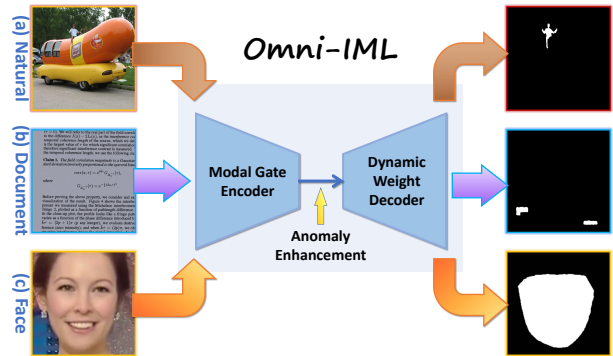


Figure 1. The proposed *Omni-IML* is the first generalist model for image manipulation localization. It that can simultaneously achieve high performance forgery localization on natural images, document images and face images with a single model, without task-specific or benchmark-specific fine-tuning.

Despite the progress made in recent years, existing IML models are designed for individual image types (e.g. natural style images, document images, face images). Although these specialized models can handle multiple tampering methods on the images of a single target type, they always fall short on other types of forged images. The lack of generality notably increases the maintenance costs of IML, since an additional image type classifier and multiple IML models must be maintained for different image types. In addition, the error accumulation caused by image type misclassification is still severe, as the existing IML models perform poorly on the image types they are not designed for. This significantly hinders the real-world application of IML. It is crucial to develop a generalist IML model that can simultaneously perform well on all image types.

Jointly training an IML model on diverse image types can slightly alleviate the random guessing issue on different image types. However, in most cases, the joint training will lead to an obvious performance degradation on all image types, making the predictions unreliable. For example, *HiFi-Net* [11] suffers from joint training and thus uses two different sets of model parameters for natural images and face images separately. There are two main reasons why

existing IML methods suffer so much from joint training:

First, existing IML methods rely heavily on specific architecture designs, input modalities, and training strategies to detect specific tampering clues on specific image types. These designs work well for the target image type, but usually not so well for other image types. For example, edge anomaly enhancement modules [6] and object attention modules [32] have made significant progress in identifying forged natural objects. However, they can hardly work well on document and face images where edge artifacts are not obvious. Early frequency-vision [26] fusion achieves satisfactory performance on document images but has obvious performance degradation on natural and face images that cover much more noise and diversity. The high-resolution representation learning design with shallow layers [11, 19] performs well in capturing the texture anomalies left by deepfake models but falls short on natural and document images where the tampered regions are small and the texture anomalies are not obvious.

Second, existing IML methods lack the design to alleviate the confusion in unified IML modeling. The IML task is already challenging since various tampering methods have already produced different unobvious tampering cues on each single image type, learning a general representation for tampering cues on different image types could be even more challenging. Without a suitable design, models will be easily confused when learning to distinguish so many tampering features from authentic ones.

To address the above issues, we propose Omni-IML, the first generalist model that can simultaneously perform well on all three major IML tasks with a single model, as shown in Fig 1. Specifically, a Modal Gate Encoder is proposed to automatically select the optimal encoding modality for each input sample, based on the characteristics of the input image. Additionally, a Dynamic Weight Decoder is proposed to adaptively select the optimal decoder filters for each sample, assisting the generalist model to better cope with the highly diverse tampering features from different tampering methods on multiple image types. These sample-adaptive designs effectively help the model achieve generalism through flexibly adapting itself to each sample. Further, an Anomaly Enhancement module is introduced between the encoder and decoder. It enhances the features of tampered regions with a novel box supervision design and suppresses the noise introduced by the joint learning on different tampering methods and image types.

We validate the effectiveness of our Omni-IML on three representative IML tasks, including natural IML, document IML, and face IML. Without bells and whistles, experimental results showcase that our single model achieves state-of-the-art performance simultaneously on all three tasks, significantly surpassing previous specialized methods on individual tasks. These strong results verify the design of our

generalist model in the field of image forensics.

By unifying the IML on natural images, document images and face images with a single model, our Omni-IML successfully eliminates the trouble of judging image type at first and maintaining different models for diverse image types. The issues of severe error accumulation and high maintenance costs are thus well solved, significantly promoting the real-world applications of IML. The development of Omni-IML is also in line with the current main trend towards Artificial General Intelligence (AGI).

In summary, our main contributions are as follows:

- We propose Omni-IML, **the first generalist model** for image manipulation localization, which serves as a pioneering effort in this field.
- Our technical innovations lie in the **novel and effective modules**: (1) Modal Gate Encoder to effectively select sample-specific encoding modality, facilitating better modality collaboration. (2) Anomaly Enhancement, which enhances the common features of the forged regions through task collaboration. (3) Dynamic Weight Decoder, which adaptively selects the sample-specific decoder filters and reduces conflicts in the unified training.
- Extensive experiments demonstrate that our generalist model can **simultaneously** achieve **state-of-the-art** results **with a single model** on natural image IML, document IML and face IML.

2. Related works

2.1. Specialized Image Manipulation Localization

Natural Image Manipulation Localization aims to identify the tampered regions in daily-life style images. MantraNet [37] proposes to perform natural IML with noise filters SRM and Bayar Conv. MVSS-Net [6] introduces ESB module to enhance boundary inconsistency. ObjectFormer [32] proposes an object encoder to learn object-level attention for better feature extraction and proposes BSCIM module to enhance the edge inconsistency. TruFor [10] benefits from the noise filters Noiseprint++. UnionFormer [28] introduces a new backbone to enhance edge artifacts, and proposes to model the inconsistency between tampered objects and authentic objects. These model designs have achieved significant progress in natural images, but their performance in document and face forensics scenarios is unsatisfactory due to the the absence of natural object, edge artifacts and noise artifacts in these scenarios.

Document Image Manipulation Localization aims to localize the forged regions in document images. Early works [3, 31] achieve document forensics through template-matching based methods. These methods work well on clean documents but do not excel on complex documents such as photographed documents, and even cannot work on natural or face images. Document Tampering Detector [26]

improves document IML through early fusion of vision and frequency features. However, the model will be seriously distorted in many cases of natural and face images where the frequency features are too noisy. TIFDM [8] proposes high-level spatial attention to suppress the false alarms in documents, but it is limited on complex natural images.

Face Image Manipulation Localization aims to localize fake human faces. The advancement of deepfake techniques makes it easy to generate a face that does not exist [5, 15]. To ensure the security of face images and improve the interpretability of deepfake detection, some recent works have explored face image forgery localization, characterized by a shallow network design for texture artifacts detection. HiFiNet [11] utilizes metric learning for better texture anomaly capturing. DA-HFNet [20] proposes Dual Attention Feature Fusion to better capture the AIGC artifacts. These methods show generalization on face IML but are sub-optimal on natural and document images, where the tampered regions are small in size and the visual anomalies are less obvious.

2.2. Generalist Model

Recently, generalist model has attracted increasing attention since it is more convenient for academic and application [33]. Despite the progress in unified object detection and segmentation [34], most of the previous generalist models do not cover all image forensic tasks. EVP [18] unifies natural image forensics with other low-level tasks such as shadow detection, but it can only perform IML on natural images and its performance is not satisfactory enough. Therefore, EVP cannot be considered as a generalist model for IML. For image forgery localization, none of the existing work realizes a unified model that can be simultaneously generalized to natural images, document images and face images. It is still unexplored towards **a generalist IML model** that can generalize on various tampering methods **across different image types**.

3. Methodology

As shown in Figure 2, the overall architecture of the proposed Omni-IML is roughly based on encoder-decoder architecture. The Modal Gate Encoder of the Omni-IML consists of four modules: (1) Visual Perception Head (VPH) to extract visual features from the original images; (2) Frequency Perception Head (FPH) to convert the Discrete Cosine Transform (DCT) coefficients of the images to frequency domain features; (3) a Modal Gate to automatically determine the optimal modality for the following encoding process; (4) a backbone model to extract multi-scale high-level features from the output of the Modal Gate. The Dynamic Weight Decoder of the Omni-IML adaptively selects the sample-specific optimal decoder filters and outputs the final mask prediction. We also design an Anomaly En-

hancement module between the encoder and decoder, to enhance the common features of tampered regions from various image types.

3.1. Modal Gate Encoder

Key Idea. The frequency feature is a double-edged sword for the IML generalist. The frequency feature can help to detect visually consistent tampering in some cases, but it can also degrade the model performance when the image is complex and noisy, or the frequency information is not prominent in the original image. As a result, neither pure vision modeling nor vision+frequency modeling can consistently provide the optimal solution. In order to achieve general IML through a flexible encoding modality, we propose the Modal Gate, which automatically determines the optimal encoding modality (frequency+vision or pure vision) for each input sample. The key idea of our Modal Gate Encoder **is to automatically identify the optimal modality by analyzing whether the frequency features contain too much noise, and which coarse prediction seems more confident, reliable, and accurate.**

Image Encoding. As shown in Figure 2, the Omni-IML considers both vision domain modeling and frequency domain modeling. Given an input image $X \in \mathbb{R}^{H \times W \times 3}$ and its Y-channel quantization table $QT \in \mathbb{R}^{8 \times 8}$, we extract vision features F_{rgb} using Visual Perception Head (VPH), $F_{rgb} = VPH(X)$. We obtain frequency features F_{freq} from the DCT coefficients and quantization tables (QT) of the images using Frequency Perception Head (FPH), $F_{freq} = FPH(DCT(X), QT)$. We use the same VPH and FPH architectures as those proposed in Document Tampering Detector [26]. The F_{freq} is fused with F_{rgb} by a channel-spatial attention module $Attn$ to get the fused features F_{fused} , $F_{fused} = Attn(F_{rgb}, F_{freq})$. Two coarse binary mask predictions P_{rgb} and P_{fused} are further obtained from F_{rgb} and F_{fused} with two auxiliary heads $AuxHead$ respectively, $P_{rgb} = AuxHead_1(F_{rgb})$, $P_{fused} = AuxHead_2(F_{fused})$, each of the auxiliary heads consists of two conv-layers.

Modal Gate. As shown in Figure 3, the input of the proposed Modal Gate has four parts: F_{rgb} , F_{fused} , P_{rgb} and P_{fused} ; We repeat P_{rgb} , P_{fused} and concatenate them with F_{rgb} , F_{fused} to get F_{cat} , which is then fed into a binary classifier for optimal modality prediction. $P_{cls} = CLS(F_{cat})$, $P_{modal} = Round(\sigma(P_{cls}))$, where σ is the sigmoid function and $Round$ is the rounding up function. The classifier CLS consists of several conv-layers, a global average pooling layer and a linear layer, and is used to determine whether to use the fused feature F_{fused} or the pure vision feature as the encoder input F_{rgb} , **by observing the noise level and anomaly significance level** of F_{fused} , F_{rgb} and their corresponding coarse predictions P_{rgb} and P_{fused} .

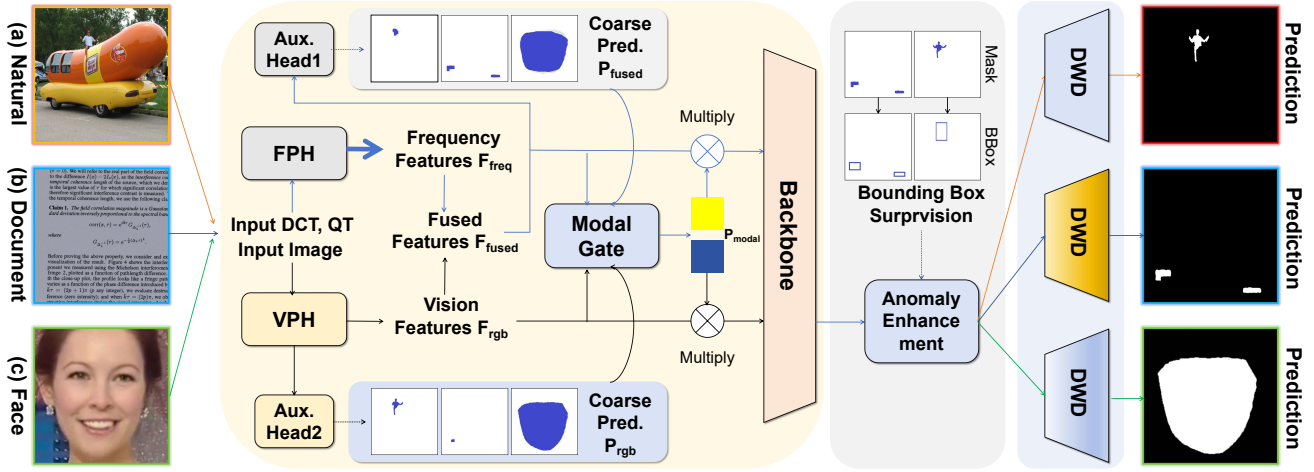


Figure 2. The overall framework of the proposed Omni-IML.

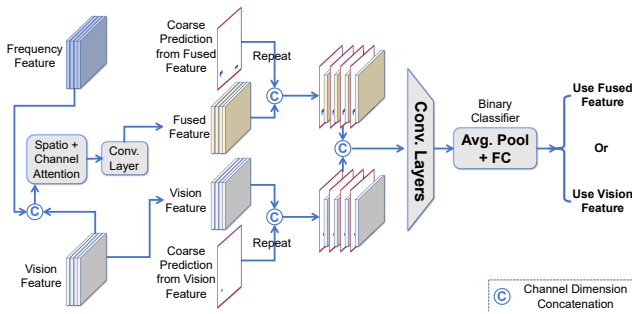


Figure 3. The proposed Modal Gate.

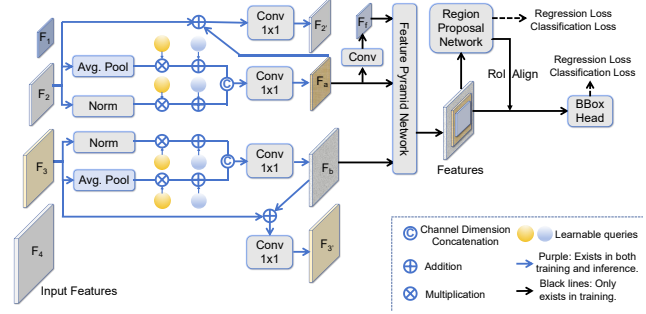


Figure 4. The proposed Anomaly Enhancement module.

Loss Function. The Modal Gate Encoder is optimized with L_{MG} , the sum of two segmentation losses and one classification loss. CE denotes the cross-entropy loss function, L_m is the ground-truth mask indicating tampered region and $L_c \in \{0, 1\}$ is the classification label indicating the optimal modality. L_c is obtained by choosing the most accurate coarse prediction. $IoU(x, y)$ denotes the Insert over Union between inputs x and y .

$$L_{MG} = CE(P_{rgb}, L_m) + CE(P_{fused}, L_m) + CE(P_{cls}, L_c)$$

$$L_c = \begin{cases} 1 & IoU(P_{rgb}, L_m) > IoU(P_{fused}, L_m) + 0.1 \\ 0 & otherwise \end{cases}$$

The Modal Gate Encoder maximizes the advantages of frequency domain modeling especially when the visual anomalies are limited (e.g. document images), and avoids its drawbacks when the image is too complex and noisy (e.g. natural images). Our Modal Gate Encoder extracts the best features from different image types and thus considerably benefits the generalist IML model.

3.2. Anomaly Enhancement

Key Idea. Sophisticated tampering leaves very obscure anomaly clues. The encoder’s output feature from such challenging sample can be very noisy. Different image types produce different features and thus joint training brings much more noise to the features and confuses IML model. To tackle this, we propose to enhance the features of forged regions and suppress the noise through including an extra box supervision during training. Since the detection framework has a clear different characteristic from the original segmentation one, training the model under both frameworks further highlights anomaly features by reducing the learning bias: If a feature region reports positive under both the detection and segmentation frameworks, it can mostly be the actual tampered region. However, if a feature region reports positive under only one framework, it is likely to be a false-positive noise and will be punished under the other framework. As a result, the contrast between the features of forged regions and authentic regions can be strengthened, noise can be suppressed and the common tampering features can be learned. However, directly training the model with the two frameworks may also cause task competition

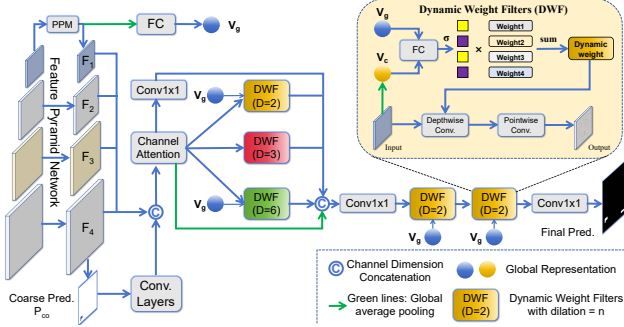


Figure 5. The proposed Dynamic Weight Decoder.

for model parameters [12] and weaken model performance, while directly scaling up the model parameters could alleviate the competition but will increase computation burden. To address this issue, we propose a novel effective collaboration module Anomaly Enhancement (AE).

Method. As shown in Figure 4, for the input features F_2 and F_3 , we first extract task-agnostic features F_a and F_b with query-based attention, the learnable attention queries contain prior knowledge to decouple and to minimize negative impact from the segmentation supervision. After that, F_a and F_b are processed by the detection modules, including two Feature Pyramid Networks (FPNs) [17] and the Faster R-CNN’s [29] RPN and RoI-Heads. The detection modules (black arrows in Figure 4) are only present during training. Including the two cascaded FPNs reduces parameter competition from the detection framework and discarding them during inference ensures the computation efficiency, successfully addressing the dilemma. After training, the F_a and F_b contain positive features enhanced by the detection supervision, we add them to the original features F_2 and F_3 and fuse them with conv-layer to get $F_{2'}$ and $F_{3'}$.

Loss Function. As shown in Figure 4, the AE module is optimized by bounding box losses as Faster R-CNN [29] from the RPN and RoI-Head. $L_{AE} = L_{cls}^{RPN} + L_{regression}^{RPN} + L_{cls}^{RoIHead} + L_{regression}^{RoIHead}$. The ground-truth boxes are the bounding boxes of the mask labels’ connected regions.

The AE module is tested in an end2end manner as shown in Figure 4. The proposed AE effectively achieves task collaboration while keeping the inference cost almost unchanged. With the proposed AE module, the tampered regions in features F_2 and F_3 can be enhanced and the false-positive noise can be reduced. Consequently, our AE module helps to extract better common features and thus benefits the generalist model.

3.3. Dynamic Weight Decoder

Key Idea. Different types of tampered image result in a wide range of manipulation clues. For example, forged objects in natural style images may have abnormal contrast or edge artifacts [32], tampered text in document images

might be visually consistent but has discontinuous BAG in frequency domain [26], fake faces may have unnatural texture [11]. These wide variations of tampering clues further cause a large variation of the encoded features of tampered regions. Merely using a fixed set of filters for the decoder causes it being confused by the diverse encoder features, especially in the unified training process. To address this challenge, we propose to adaptively select the optimal decoder filters for each input image based on the characteristics of the image and the initial predicted tampered region. To achieve this, we propose the DWD, as shown in Figure 5.

Method. In the proposed Dynamic Weight Decoder, the low-level input features are fused with high-level input features by Pyramid Pooling Module [39] and Feature Pyramid Network [17] to obtain multi scale features F_1, F_2, F_3, F_4 . A global feature vector V_g is obtained by average pooling F_4 . A coarse mask prediction P_{co} is obtained from the lowest-level feature F_1 by a conv-layer, $P_{co} = Conv(F_1)$. A light-weight network CNN is used to extract features F_{co} from the coarse prediction P_{co} , $F_{co} = CNN(P_{co})$. The extracted feature is concatenated to the multi-scale features and it helps the model to pay attention to the suspicious regions and analyze the forgery type, $F_{cat} = Concat(F_1, F_2, F_3, F_4, F_{co})$. The concatenated features are channel dimension reduced and processed by a series of Dynamic Weight Filters (DWF) with different dilation rates, $F_{dec1} = Concat(Avg(F_{cat}), F_{dw}, F_{cat})$, $F_{dw} = Concat([DWF_n(F_{cat}, V_g) \text{ for } n \text{ in } (2, 3, 6)])$, DWF_n denotes the proposed DWF with dilation rate n . The final prediction P_{DWD} is obtained by $P_{DWD} = Conv(DWD_2(DWD_2(Conv(F_{dec1}), V_g), V_g))$, where $Conv$ denotes 1×1 conv-layer. The DWD is surprised by minimizing the cross-entropy loss between P_{DWD} , P_{co} and the ground-truth mask L_m . $L_{DWD} = CE(P_{DWD}, L_m) + CE(P_{co}, L_m)$

Dynamic Weight Filters. As shown in the top-right of Figure 5, to obtain the dynamic filters, we first average pool the input feature to obtain a current global representation V_c (orange box in Figure 5), then interact V_c with the global image vector V_g (blue box in Figure 5) with a fully connected layer and identify the optimal dynamic filters D_{opt} by weighted summation of four common convolutional filters. $A_i = \sigma(FC(V_c, V_g))$, $D_{opt} = \sum_{i=1}^4 A_i * W_i$, σ is the sigmoid function, FC is the linear layer, W_i is the i th filter in the DWF. Finally, we depth-wise convolve the input feature with D_{opt} and then perform point-wise convolution with 1×1 conv-layer to obtain the output.

The proposed DWD achieves sample-specific filters selection by analyzing the characteristics of the input image, the input features and the forgery types in the initially predicted tampered region. The selected optimal filters effectively help the generalist model to simultaneously distinguish tampered regions in different image types.

Table 1. Comparison study on natural image manipulation localization. The training data of 'CAT-Netv2' and 'TruFor' includes the entire IMD20 dataset, thus their performance on IMD20 is not evaluated.

| Method | Omni | CASIaV1 | | Coverage | | CocoGlide | | NIST16 | | IMD20 | | Avg. (w.o. IMD) | | Avg. (w/ IMD) | |
|-----------|------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-----------------|-------------|---------------|-------------|
| | | IoU | F1 | IoU | F1 | IoU | F1 | IoU | F1 | IoU | F1 | mIoU | mF1 | mIoU | mF1 |
| ManTraNet | No | .086 | .130 | .181 | .271 | .310 | .408 | .040 | .062 | .098 | .146 | .154 | .218 | .143 | .203 |
| RRU-Net | No | .330 | .380 | .165 | .260 | .223 | .304 | .080 | .129 | .169 | .256 | .200 | .268 | .193 | .266 |
| MVSS-Net | No | .403 | .455 | .389 | .454 | .278 | .360 | .243 | .294 | .243 | .294 | .328 | .391 | .311 | .371 |
| PSCC-Net | No | .410 | .463 | .340 | .446 | .333 | .422 | .067 | .110 | .115 | .192 | .288 | .360 | .253 | .327 |
| CAT-Netv2 | No | .684 | .738 | .238 | .292 | .290 | .366 | .238 | .302 | - | - | .363 | .425 | - | - |
| IF-OSN | No | .465 | .509 | .181 | .268 | .259 | .364 | .247 | .326 | .259 | .364 | .288 | .367 | .282 | .366 |
| EVP | No | .438 | .502 | .078 | .114 | .232 | .346 | .188 | .239 | .177 | .268 | .234 | .300 | .223 | .294 |
| TruFor | No | .630 | .692 | .446 | .522 | .294 | .362 | .279 | .348 | - | - | .412 | .481 | - | - |
| APSC-Net | No | .810 | .848 | .498 | .568 | .392 | .455 | .525 | .590 | .679 | .760 | .556 | .615 | .581 | .644 |
| Ours | Yes | .798 | .834 | .524 | .576 | .448 | .505 | .556 | .630 | .662 | .740 | .582 | .636 | .598 | .657 |

Table 2. Comparison study on document image manipulation localization.

| Method | Omni | SACP | | DocTamper-TestingSet | | | | DocTamper-FCD | | | | DocTamper-SCD | | | |
|-----------------|------------|-------------|-------------|----------------------|-------------|-------------|-------------|---------------|-------------|-------------|-------------|---------------|-------------|-------------|-------------|
| | | IoU | F1 | IoU | P | R | F1 | IoU | P | R | F1 | IoU | P | R | F1 |
| DFCN [41] | No | .466 | .607 | - | - | - | - | - | - | - | - | - | - | - | - |
| MVSS-Net [6] | No | .401 | .534 | - | - | - | - | - | - | - | - | - | - | - | |
| SE-Net [38] | No | .459 | .587 | - | - | - | - | - | - | - | - | - | - | - | |
| RRU-Net [4] | No | .517 | .651 | - | - | - | - | - | - | - | - | - | - | - | |
| CFL-Net [24] | No | .433 | .571 | - | - | - | - | - | - | - | - | - | - | - | |
| TIFDM [8] | No | .576 | .703 | - | - | - | - | - | - | - | - | - | - | - | |
| ManTraNet [37] | No | - | - | .180 | .123 | .204 | .153 | .170 | .175 | .261 | .209 | .160 | .124 | .218 | .157 |
| MVSS-Net [6] | No | - | - | .430 | .494 | .383 | .431 | .410 | .480 | .381 | .424 | .400 | .478 | .366 | .414 |
| PSCC-Net [19] | No | - | - | .170 | .309 | .506 | .384 | .160 | .440 | .580 | .420 | .190 | .286 | .540 | .374 |
| BEiT-Uper [2] | No | - | - | .590 | .564 | .451 | .501 | .350 | .550 | .436 | .487 | .340 | .408 | .395 | .402 |
| Swin-Uper [21] | No | - | - | .700 | .671 | .608 | .638 | .410 | .642 | .475 | .546 | .510 | .541 | .612 | .574 |
| CAT-Netv2 [14] | No | - | - | .710 | .768 | .680 | .721 | .600 | .795 | .695 | .741 | .540 | .674 | .665 | .670 |
| DTD [26] | No | - | - | .828 | .814 | .771 | .792 | .749 | .849 | .786 | .816 | .691 | .745 | .762 | .754 |
| Omni-IML (Ours) | Yes | .714 | .820 | .842 | .837 | .802 | .819 | .750 | .901 | .760 | .824 | .685 | .760 | .786 | .773 |

Table 3. Comparison study on models trained on all tasks.

| Method | Official model trained on specific tasks | | | | | | | | Re-trained on all tasks with the same settings | | | | | | | |
|----------|--|------|------|------|-----------|------|------|------|--|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | Natural | | SACP | | DocTamper | | Face | | Natural | | Document | | Face | | Avg. | |
| | mIoU | mF1 | mIoU | mF1 | mIoU | mF1 | mIoU | mF1 | mIoU | mF1 | mIoU | mF1 | mIoU | mF1 | mIoU | mF1 |
| EVP | .223 | .294 | .030 | .053 | .016 | .035 | .305 | .453 | .455 | .501 | .411 | .447 | .814 | .886 | .560 | .611 |
| HiFi-Net | .023 | .032 | .106 | .116 | .078 | .109 | .784 | .815 | .447 | .492 | .427 | .461 | .815 | .892 | .563 | .615 |
| DTD | .037 | .059 | .140 | .224 | .756 | .787 | .003 | .005 | .314 | .372 | .468 | .501 | .820 | .901 | .534 | .591 |
| TIFDM | - | - | .576 | .703 | - | - | - | - | .473 | .515 | .432 | .473 | .820 | .900 | .575 | .629 |
| APSC-Net | .581 | .644 | .088 | .133 | .139 | .184 | .151 | .197 | .587 | .653 | .616 | .657 | .818 | .900 | .674 | .737 |
| Ours | - | - | - | - | - | - | - | - | .598 | .657 | .748 | .809 | .822 | .902 | .723 | .789 |

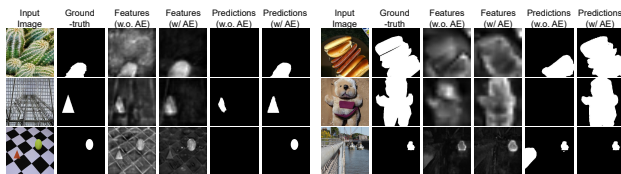


Figure 6. Visualization for the ablation of the AE module.

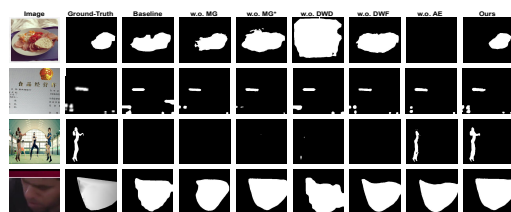


Figure 7. Qualitative results for visual comparison.

4. Experiments

4.1. Experiment Setup

Training Data. The training data includes three parts:

(1) Natural style image. We utilize the tampCOCO [14], CASIAv2 [7], MIML [28], and COCO [16] datasets as the training set of the natural image part, following the standard practice in the IML field [10, 28].

(2) Document image. SACP [1] and DocTamper [26] are high-quality, large-scale document IML datasets with varied tampering methods. We include the training sets of SACP and DocTamper as the document image part.

(3) Face image. We use the training set of the FaceShifter subset from HiFi-IFDL [11] and 24k random images from CelebHQ [13] as the face image part.

Test Data. The test data of Omni-IML includes three parts:

(1) Natural style image. We adopt the widely used benchmarks CASIAv1 [7], Coverage [35], NIST16 [9] and IMD20 [25] for evaluation. These benchmarks include diverse tampered objects of various styles and diverse handcrafted forgeries of various types (e.g. copy-move, splicing, removal). We also include the CocoGlide dataset [10] which contains forgeries produced by diffusion model.

(2) Document image. We use the test set of SACP [1], which contains handcrafted forgeries of various types (e.g. copy-move, splicing, removal, printing, AIGC-based editing) and heavy post-processing. We also include the three test sets from the DocTamper benchmark [26], which contains high-quality forgeries and can evaluate IML models in both in-domain and out-of-domain scenarios.

(3) Face image. The FaceShifter test set [11] is adopted as the face image part. These fake faces are produced by the representative DeepFake model FaceShifter [15].

Implementation Details. The backbone model of our Omni-IML is ConvNeXt-Base [22] initialized with its official ADE20k [40] pre-trained weights, following previous works [10, 28]. The Omni-IML is trained with the cross-entropy loss for 370k iterations, using the AdamW optimizer [23], with a batch size of 16 and an input size of 512×512 . The initial learning rate is set to $1e-4$ and decays to $1e-6$ in a linear schedule. A fixed threshold of 0.5 is used to binarize model predictions during inference.

Evaluation Metrics. For the DocTamper benchmark, we use the official scripts to evaluate model performance. For other benchmarks, we calculate fore-ground IoU and pixel-level Precision (P), Recall (R), and F1-score (F) for each sample and then compute the average score following the previous work [28] for fair comparison.

4.2. Comparison Study

The proposed generalist model Omni-IML is evaluated on all of the natural IML, document IML, and face IML benchmarks using a single set of model parameters, without any

task-specific or benchmark-specific fine-tuning. The comparison with the state-of-the-art methods of natural image forensics is shown in Table 1, the methods compared include Mantra-Net [37], RRU-Net [4], MVSS-Net [6], PSCC-Net [19], CAT-Netv2 [14], IF-OSN [36], EVP [18], TruFor [10], APSC-Net [28]. The comparison with the state-of-the-art methods for document IML and face IML tasks are shown in Table 2 and Table 4, respectively. Evidently, our generalist Omni-IML can simultaneously outperform existing specialized methods on each individual task, demonstrating the strong generalization ability. This is because our Omni-IML can adaptively select the optimal input modality and decoder parameters for each sample, effectively producing the best features for IML on different image types. In addition, the Anomaly Enhancement module drives the model to learn common features for the forgeries from different image types, and reveals the inconsistencies between forged and authentic regions with the extra box supervision. Consequently, it suppresses feature noise and reduces model confusion in joint training.

It’s worth noting that in Table 4, the HiFi-Net provides two separate official models for IML on natural images and face images respectively. This is because the HiFi-Net suffers greatly from joint training, and it is necessary to train it separately for each task. Furthermore, HiFi-Net and TruFor only perform well with their specialist face IML models, while our Omni-IML excels with a generalist model, demonstrating the effectiveness of our methods.

To further explore the generalist capability of previous IML methods, we re-train the state-of-the-art models with their official model code, the same training data and pipeline as ours, the results are shown in Table 3. In Table 3, the left part is the performance of their official model trained on specific tasks. Evidently, all the models perform well on only one task. The right part of Table 3 is the performance of the re-trained models. The average performance of the re-trained models improves as joint training alleviates the random guessing issue on other image types. Including the MIML dataset for training also counteracts the significant performance degradation brought by joint training on diverse image types. Despite this, they still perform significantly worse than our Omni-IML (e.g. 5-20 points mIoU lower than ours). This is because existing IML methods rely heavily on designs and strategies targeted at one image type, and such designs and strategies usually do not work so well on other image types (e.g. noise filters, edge enhancement and object-level attention are beneficial for natural images but not for document images). Moreover, the tampering features among diverse image types differ a lot from each other, making it challenging for models to simultaneously learn them well. As a result, training IML models jointly on image types for which they are not designed causes considerable confusion and significantly lim-

Table 4. Comparison study on face forgery localization. ‘Face re-trained’ denotes the model re-trained on the FaceShifter data using official code. ‘Natural model’ and ‘Face model’ denote the official models trained on natural images and face images respectively.

| Method | Omni | IoU | P | R | F1 |
|-------------------------------|------------|-------------|-------------|-------------|-------------|
| TruFor (Official model) [10] | No | .631 | .984 | .638 | .774 |
| TruFor (Face re-trained) [10] | No | .814 | .990 | .819 | .896 |
| HiFi-Net (Natural model) [11] | No | .255 | .439 | .379 | .407 |
| HiFi-Net (Face model) [11] | No | .784 | .866 | .800 | .815 |
| Omni-IML (Ours) | Yes | .822 | .993 | .826 | .902 |

Table 5. Ablation study on the proposed modules.

| Ablation | Natural | | Document | | Face | | Average | |
|----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | mIoU | mF1 | mIoU | mF1 | mIoU | mF1 | mIoU | mF1 |
| Baseline | .451 | .544 | .509 | .580 | .809 | .888 | .589 | .670 |
| w.o. MG | .500 | .552 | .609 | .672 | .810 | .890 | .639 | .704 |
| w.o. MG* | .568 | .632 | .625 | .673 | .811 | .889 | .668 | .731 |
| w.o. DWD | .477 | .567 | .515 | .580 | .815 | .894 | .602 | .680 |
| w.o. DW | .562 | .625 | .692 | .765 | .820 | .901 | .691 | .763 |
| w.o. AE | .547 | .601 | .662 | .726 | .819 | .900 | .676 | .742 |
| Ours | .598 | .657 | .748 | .809 | .822 | .902 | .723 | .789 |

Table 6. Ablation study on the backbone model size.

| Backbone | Natural | | Document | | Face | | Average | |
|----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | mIoU | mF1 | mIoU | mF1 | mIoU | mF1 | mIoU | mF1 |
| ConvNeXt Small | .588 | .648 | .736 | .793 | .821 | .901 | .715 | .781 |
| ConvNeXt Base | .598 | .657 | .748 | .809 | .822 | .902 | .723 | .789 |
| ConvNeXt Large | .605 | .665 | .770 | .829 | .836 | .910 | .737 | .801 |

its their performance. Our Omni-IML does not rely on modules or strategies that designed for only one image type. In contrast, the adaptive selection of optimal encoding modality and decoder parameters helps our model to effectively handle diverse tampering clues and extract the best features from various image types. Additionally, the anomaly enhancement also benefits all domains by enhancing the features of tampered regions and driving the model to learn common features from diverse image types. Consequently, our Omni-IML demonstrates strong generalization across different image types and has minimal performance degradation during joint training.

Ablation Study on the Proposed Modules. The ablation results are shown in Table 5. ‘w.o. MG’ denotes the model without the Modal Gate, it has 8.4 points lower mIoU than Omni-IML. This is because the frequency features in some samples are unstable, and without the Modal Gate to filter them out, these features introduce too much noise to the encoder and thus cause performance degradation. ‘w.o. MG*’ represents the model without Modal Gate and using the pure

vision modality, it has 5.5 points lower mIoU than Omni-IML. This is because frequency domain modeling can also be helpful in some cases, especially when the tampered region is visually consistent (e.g. on document images). ‘w.o. DWD’ represents the model without the Dynamic Weight Decoder, it has 12.1 points lower mIoU than Omni-IML. This is because the diversity of tampering features is too high for the encoder to learn them well, thus confusing the model, confirming the necessity of the proposed DWD for the generalist model. ‘w.o. DW’ is the model with the DWD structure but the filter weights in the decoder keep all the same for each input, it has 3.2 points lower mIoU than Omni-IML, this verifies that the adaptive selection of optimal decoder weights for each sample can reduce confusion in joint training. ‘w.o. AE’ is the model without the proposed Anomaly Enhancement (AE) module, it has 4.7 points lower mIoU than Omni-IML. This is because the proposed AE module can enhance the forged regions in the features, and can drive the model to learn common features. Without the AE module, the encoder’s output features will have much more noise and confuse the decoder, as visualized in Figure 6, The model without any of the proposed modules serves as the ‘Baseline’ model, its mIoU is 13.4 points lower than Omni-IML. These results have proved the effectiveness of our methods.

Ablation Study on the Model Size. We conduct an ablation study on the model size. As shown in Table 6, the model performance improves slightly with a larger size. These results indicate the scaling law behind our Omni-IML and there is a great potential for further improvement.

5. Conclusion

In this paper, we propose Omni-IML, the first generalist model designed for image manipulation localization to address the drawbacks of specialist models. Specifically, multiple novel and effective modules are proposed to achieve generalism through sample-specific adaptation, including a Modal Gate Encoder that automatically determines the optimal encoding modality for each input image, and a Dynamic Weight Decoder that adaptively selects the optimal decoder parameters for each input sample. In addition, an Anomaly Enhancement module is proposed to reduce confusion by enhancing the features of tampered regions and driving the model to learn common features from diverse image types. To verify the generalist capability, extensive experiments are conducted on three major IML tasks, covering natural IML, document IML, and face IML. The experimental results demonstrate that our single model simultaneously achieves state-of-the-art performance on all tasks. Comprehensive ablation studies and visual analyses are also presented to provide in-depth insights. We believe that our work can inspire future research and promote the real-world applications of unified image forensics models.

References

- [1] Alibaba Security. Security ai challenger program. <https://tianchi.aliyun.com/competition/entrance/531812/introduction>, 2020. 7
- [2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 6
- [3] Bilal Bataineh, Siti Norul Huda Sheikh Abdullah, and Khairudin Omar. A statistical global feature extraction method for optical font recognition. In *Intelligent Information and Database Systems: Third International Conference, ACIIDS 2011, Daegu, Korea, April 20-22, 2011, Proceedings, Part I 3*, pages 257–267. Springer, 2011. 2
- [4] 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. 6, 7
- [5] Dongyue Chen, Qiusheng Chen, Jianjun Wu, Xiaosheng Yu, and Tong Jia. Face swapping: realistic image synthesis based on facial landmarks alignment. *Mathematical Problems in Engineering*, 2019(1):8902701, 2019. 3
- [6] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022. 2, 6, 7
- [7] 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. 7
- [8] Renshuai Liu Dong, Li, Bowen Ma, Wei Zhang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, and Xuan Cheng. Robust text image tampering localization via forgery traces enhancement and multiscale attention. *IEEE Transactions on Consumer Electronics*, 2024. 3, 6
- [9] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhan, 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. 7
- [10] 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. 2, 7, 8
- [11] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *CVPR*, 2023. 1, 2, 3, 5, 7, 8
- [12] Falk Heuer, Sven Mantowsky, Saqib Bukhari, and Georg Schneider. Multitask-centernet (mcn): Efficient and diverse multitask learning using an anchor free approach. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 997–1005, 2021. 5
- [13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 7
- [14] 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. 6, 7
- [15] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2020. 3, 7
- [16] 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
- [17] 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. 5
- [18] 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. 3, 7
- [19] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscn-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. 2, 6, 7
- [20] Yang Liu, Xiaofei Li, Jun Zhang, Shengze Hu, and Jun Lei. Da-hfnet: Progressive fine-grained forgery image detection and localization based on dual attention. *arXiv preprint arXiv:2406.01489*, 2024. 3
- [21] 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. 6
- [22] 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. 7
- [23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7
- [24] Renshuai Liu Niloy, Fahim Faisal, Bowen Ma, Wei Zhang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, and Xuan Cheng. Cfl-net: image forgery localization using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4642–4651, 2023. 6
- [25] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detect-

- ing manipulated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2020. 7
- [26] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: new dataset and new solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5937–5946, 2023. 1, 2, 3, 5, 6, 7
- [27] Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Generalized tampered scene text detection in the era of generative ai, 2024. 1
- [28] Chenfan Qu, Yiwu Zhong, Chongyu Liu, Guitao Xu, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards modern image manipulation localization: A large-scale dataset and novel methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10781–10790, 2024. 2, 7
- [29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 5
- [30] Zhihao Sun, Haoran Jiang, Danding Wang, Xirong Li, and Juan Cao. Saff-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. 1
- [31] Joost Van Beusekom, Faisal Shafait, and Thomas M Breuel. Text-line examination for document forgery detection. *International Journal on Document Analysis and Recognition (IJDAR)*, 16:189–207, 2013. 2
- [32] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022. 2, 5
- [33] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022. 3
- [34] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1130–1140, 2023. 3
- [35] 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. 7
- [36] 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. 7
- [37] Yue Wu, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019. 2, 6, 7
- [38] Yulan Zhang, Guopu Zhu, Ligang Wu, Sam Kwong, Hongli Zhang, and Yicong Zhou. Multi-task se-network for image splicing localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4828–4840, 2021. 6
- [39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 5
- [40] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 7
- [41] Peiyu Zhuang, Haodong Li, Shunquan Tan, Bin Li, and Jiwu Huang. Image tampering localization using a dense fully convolutional network. *IEEE Transactions on Information Forensics and Security*, 16:2986–2999, 2021. 6