Fusion Transformer with Object Mask Guidance for Image Forgery Analysis

Dimitrios Karageorgiou[1] Giorgos Kordopatis-Zilos[2] Symeon Papadopoulos[1]
[1]Information Technologies Institute, CERTH
[2]VRG, FEE, Czech Technical University in Prague
{dkarageo,papadop}@iti.gr kordogeo@fel.cvut.cz

Abstract

In this work, we introduce OMG-Fuser, a fusion transformer-based network designed to extract information from various forensic signals to enable robust image forgery detection and localization. Our approach can operate with an arbitrary number of forensic signals and leverages object information for their analysis – unlike previous methods that rely on fusion schemes with few signals and often disregard image semantics. To this end, we design a forensic signal stream composed of a transformer guided by an object attention mechanism, associating patches that depict the same objects. In that way, we incorporate object-level information from the image. Each forensic signal is processed by a different stream that adapts to its peculiarities. A token fusion transformer efficiently aggregates the outputs of an arbitrary number of network streams and generates a fused representation for each image patch. We assess two fusion variants on top of the proposed approach: (i) score-level fusion that fuses the outputs of multiple image forensics algorithms and (ii) feature-level fusion that fuses low-level forensic traces directly. Both variants exceed state-of-the-art performance on seven datasets for image forgery detection and localization, with a relative average improvement of 12.1% and 20.4% in terms of F1. Our model is robust against traditional and novel forgery attacks and can be expanded with new signals without training from scratch. Our code is publicly available at: https://github.com/mever-team/omgfuser

1. Introduction

Digital image forgery [32] is increasingly becoming more accessible and efficient due to the pervasive availability of sophisticated image editing algorithms as part of free or low-cost image editing applications for desktops and smartphones. Notably, multiple works [47, 53] have found that the ability of humans to detect forged images hardly exceeds the performance of random guessing, especially when the forgery is of decent or high quality. Despite the significant advances in the field of image forensics [32, 57, 64], existing detection methods are greatly challenged when tested in the wild without strong assumptions about the processing history of an image [61, 64]. To this end, the main focus of our work is to robustly detect the forged regions within images, along with an overall decision for the image, formally defined as *image forgery localization and detection* respectively [67].

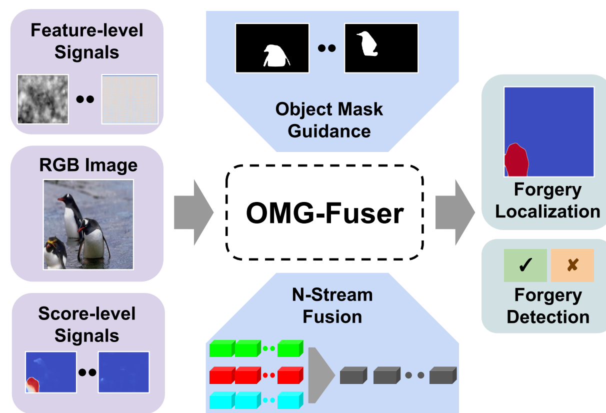


Figure 1. OMG-Fuser combines an arbitrary number of heterogeneous forensic signals for robust image forgery analysis guided by the image semantics.

A common practice for image forensics analysts is to utilize various image forensics methods to increase the chances of detecting forgery. However, in such cases, analysts need to judge based on their experience what tool to trust, an issue further exacerbated by the recent deep learning approaches that operate as black boxes. This has given rise to approaches that fuse multiple forensic signals to capture more robust forensic clues. Such approaches can be classified into two main categories [50]. (i) *Feature-level fusion*, which fuses low-level features extracted from the input image that usually represent different domains [24, 37, 68]. However, these purpose-built feature fusion architectures are only effective in detecting specific types of forgery. (ii) *Score-level fusion* [50], which combines the outputs of multiple different forensics methods into a single output [7, 20, 50]. However, theoretical limitations or the employed training approaches prevent current score-level fusion approaches from effectively combining the outputs of recent deep learning algorithms. Additionally, while image semantics could greatly help the fusion process, the impact of such a direction is underexplored in the literature [58].

To this end, we propose the *Object Mask-Guided Fusion Transformer (OMG-Fuser)*, capable of capturing image forensic traces from an arbitrary number of input signals by leveraging the image semantics and fusing them in multiple processing stages into a single robust forgery detection and localization output. Our fusion network comprises two main modules: the *Forensic Signal Streams* and the *Token-Fusion Module*. In the former, every input signal is propagated through a different stream to capture its unique traces. Our main novelty in this stage includes the *Object Guided Attention* mechanism that exploits external instance segmentation masks to drive the extraction process to attend only to image regions that depict the same objects so as to generate comprehensive object-level representations. Any pretrained instance segmentation model can be employed [22], and we utilize a recently proposed class-agnostic model [33]. The latter module is responsible for the fusion of any arbitrary number of forensic signals. The *Token-Fusion Transformer (TFT)* combines the various representations of an image region generated by the forensic signal streams into a single representation via several transformer blocks. The fused outputs are analysed by the *Long-range Dependencies Transformer (LDT)* that captures the relations between the representations of the different image regions. For training the proposed network, we propose the *Stream Drop* augmentation. This randomly discards some network streams during training to prevent the network from over-attending on specific forensic signals. Our proposed approach is employed both for feature and score level fusion, combining RGB information with two and five forensic signals, respectively. We demonstrate its effectiveness on five popular datasets, compared with several handcrafted and deep learning state-of-the-art approaches. Moreover, we demonstrate its robustness against common perturbations and recent neural filters. An overview of the approach is presented in Fig. 1.

In summary, our contributions include the following:

- We propose a fusion transformer for robust image forgery detection and localization that analyzes an arbitrary number of image forensic signals based on image semantics.
- We introduce the object guided attention that uses object-level information to drive the attention process.
- We design a token-fusion transformer that combines an arbitrary number of patch tokens into a single comprehensive representation for each image region.
- We introduce the stochastic augmentation process, named stream drop, for avoiding over-attending on particular streams while training multi-stream networks.
- We improve the state-of-the-art by 12.1% and 20.4% F1 on image forgery detection and localization, respectively.

2. Related Work

2.1. Image Forgery Detection and Localization

Image editing operations introduce subtle but detectable traces [57]. Most early works in the field of image forensics focused on detecting traces of a single type of forgery using handcrafted signal processing operations. These include, for instance, disturbances related to the Color Filter Array (CFA) [4, 17, 51] or noise inconsistencies using the popular PRNU pattern [9, 43, 66]. Others employ filtering and frequency analysis [11, 21] or extract and analyze noise residuals [44]. Also, many works focus on detecting artifacts introduced by lossy compression algorithms, such as JPEG [6, 16, 31, 39, 48].

Recent advances in deep learning have reshaped the field of image forensics. Instead of handcrafted features that capture a narrow range of artifacts and are prone to naive post-processing operations, recent methods employ deep neural networks that learn to capture more robust forensic traces for detecting forged regions. Such methods [10, 24, 29, 37] dominate the state-of-the-art, outperforming the previous approaches usually by a large margin. Several CNN-based architectures have been proposed, targeting one [2, 15, 41] or more [10, 52, 65] types of forgery, while more recently, LSTM-based [5] and transformer-based [24, 42, 55] architectures have emerged. Even though deep learning approaches can capture more complex artifacts, the employed network architectures and training procedures still limit their detection capability to a narrow range of forensic artifacts, *e.g.* focusing only on noise anomalies [28, 62] or capturing JPEG compression artifacts [37].

2.2. Fusion Approaches

According to Phan-Ho *et al*. [50], fusion approaches in the image forensics literature can be classified into two broad categories: feature-level and score-level fusion.

Feature-level fusion approaches combine feature representations incorporating information from different domains into a comprehensive one that can be employed for detecting a broader range of forgery cases. A seminal work in the field by Zhou et al. [68] proposed a two-stream network architecture: one network stream acts on the RGB image to detect forgery artifacts, and the other captures noise-related information by processing the output of SRM filters. Many state-of-the-art methods have employed such a feature fusion approach, each using different inputs and building purpose-specific fusion architectures for better capturing traces such as compression-related artifacts [37, 58], noise disturbances [12, 24] or anomalous edges [12]. However, while current feature fusion approaches significantly boost detection performance in the target forgery type, they often fail when deployed in the wild [61, 64] and are designed for fusing a small number of specific input signals.

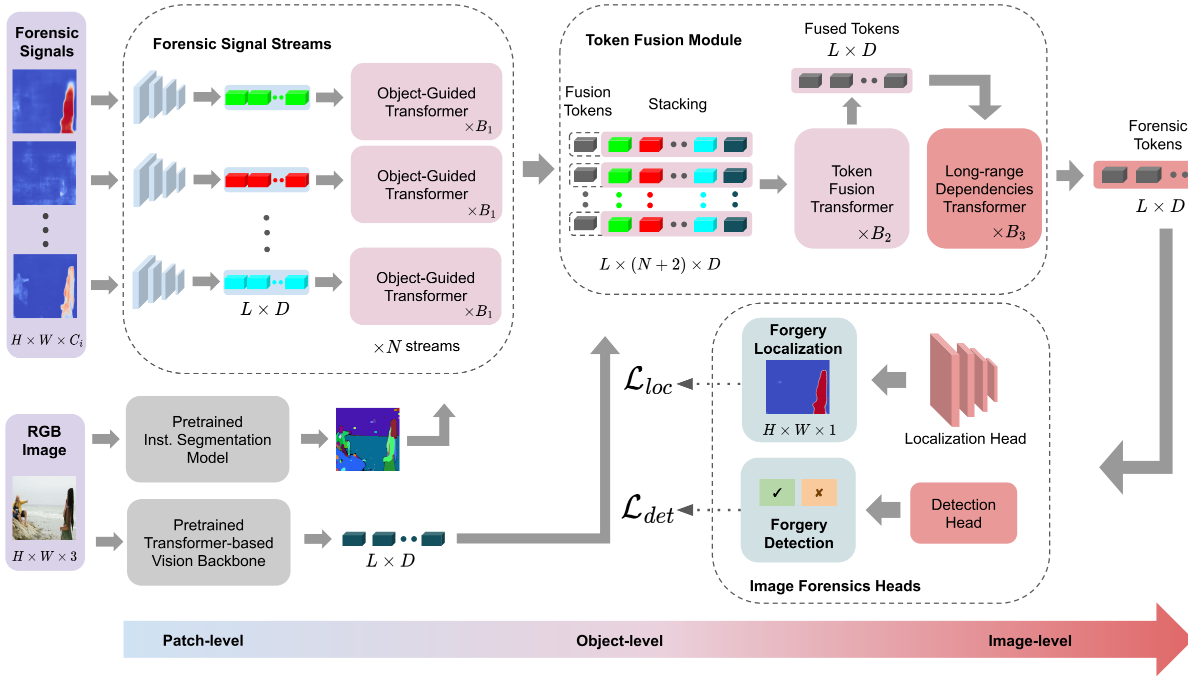


Figure 2. Overview of OMG-Fuser. Forensic signals are fused into a robust forgery localization mask and detection score. To achieve that, it combines information from the RGB image and its instance segmentation maps. Each forensic signal and the RGB image are first processed by separate network streams through independent Object-Guided Transformers. Then, the proposed Token Fusion Module fuses the different streams, leading to features with a progressively increasing level of information granularity, from patch-level (in the early stages) to object-level (in intermediate stages) and to image-level (in the final stages). A localization and a detection head process the extracted forensic tokens to generate the final outputs.

Score-level fusion approaches combine the outputs of multiple forgery detection and localization algorithms to leverage the benefits of each one into a single output. A widely adopted practice in the field is to utilize statistical approaches, such as the Dempster-Shafer Theory (DST) [18, 20] or the Bayes' Theorem [34, 35] to derive a unified prediction from multiple outputs of different image forensics algorithms. However, these require strong theoretical guarantees regarding the compatibility relations and statistical independence between the fused methods, requirements that are very tough to meet in recent deep-learning networks that operate as black boxes. Hence, they are not compatible with the current state-of-the-art algorithms. As a more general solution, learning approaches have recently been proposed for fusing the outputs of multiple image forensic algorithms [7, 19, 54]. However, they still rely on forensic signals generated by handcrafted signal processing operations as input and lack mechanisms for effectively leveraging image semantics and for preventing over-attendance on the best signals, all of which make them unsuitable for fusing state-of-the-art deep learning-based forensic signals.

Finally, recent works [58, 69] highlight the beneficial impact of utilizing semantic relations within images for forgery detection. They design trainable network components tailored to capture information related to image semantics. However, learning to incorporate information about all the possible objects encountered in-the-wild into such a network demands huge amounts of training data. To the best of our knowledge, we are the first in the field of image forensics to employ a pre-trained instance segmentation model to introduce object-level information into the attention mechanism, which requires no further training on large-scale datasets. Furthermore, there is no similar prior work in the image forensics domain that builds a transformer-based architecture that can address feature-level and score-level fusion – prior related works [7, 24, 54, 58, 68] focus on one of the two categories, and usually are limited in terms of the number of input signals that can be combined.

3. Approach Overview

This section presents our proposed **O**bject **M**ask **G**uided **Fus**ion Transform**er** (OMG-Fuser), a network architecture for fusing multiple image forensic signals by leveraging image semantics. Fig. 2 illustrates the proposed architecture.

3.1. Problem formulation

Given an RGB image $x^{rgb} \in \mathbb{R}^{H \times W \times 3}$ and a number of N forensic signals $x_i^{sig} \in \mathbb{R}^{H \times W \times C_i}, i \in \{1...N\}$, the goal is the OMG-Fuser network to predict a pixel-level forgery localization mask $\hat{y}^{loc} \in (0,1)^{H \times W \times 1}$ and an image-level forgery detection score $\hat{y}^{det} \in (0,1)$. H and W denote the height and width of the input image, respectively, while C_i denotes the number of channels of the i^{th} forensic signal.

3.2. Forensic Signal Streams

Each of the N forensic signals x_i^{sig} capture different artifact types. To capture their characteristic elements, we pass each signal through a different network stream, denoted as *Forensic Signal Stream (FSS)*. It consists of two main components, i.e., patch representations and an object-guided transformer.

Patch representations: Given that FSS is built upon transformer blocks, we need to convert the 2D input signals to a set of tokens. Following the common practice in the field [14, 58, 63], a convolutional neural network (CNN) with four strided layers is applied to the input. Its output $\bar{x}_i^{sig} \in R^{H' \times W' \times D}$ consists of feature representations extracted from the input forensic signals corresponding to specific image patches. $H' = H/p$ and $W' = W/p$ are the output's spatial dimensions, p denotes the size of the patches in the initial image, and D is the latent dimensionality. Finally, we reshape $\bar{x}_i^{sig}$ from $R^{H' \times W' \times D}$ to $R^{L \times D}$, where $L = H'W'$, to form our token sequence, which is our patch-level representations.

Object-Guided Transformer (OGT): This is the key component to infuse object-level information in the patch representations, effectively converting them to object-level representations. Inspired by Vision Transformer (ViT) [14] and Masked-Attention [8], we design a process that utilizes instance segmentation maps generated by an instance segmentation model to restrict the transformer's attention to tokens belonging to the same objects.

Given an input image, a number of 2D instance segmentation maps $x_j^{seg} \in \{0, 1\}^{H \times W \times 1}, j \in \{1...K\}$ are generated corresponding to a set of objects $\mathcal{S} = \{o_1, o_2, ..., o_K\}$ depicted in the image. Using these instance segmentation maps, we can extract a subset $\mathcal{S}_u \subset \mathcal{S}$ for each image patch u, containing all objects depicted in the corresponding $p \times p$ area in the original image. An object belongs to a patch if at least one pixel of the patch has been annotated with the corresponding object label. Hence, we define the object-guided attention mask $M \in \{-\infty, 0\}^{L \times L}$ based on the patches that contain the same objects and in order to be used during the attention calculation in the next operation. The mask generation can be formulated as

$$M_{(u,v)} = \begin{cases} 0 & \text{if } \mathcal{S}_u \cap \mathcal{S}_v \neq \emptyset \\ -\infty & \text{otherwise} \end{cases}, \tag{1}$$

where $u, v \in \{1...L\}$ denote two arbitrary image patches. In that way, during the attention calculation, image patches that depict the same objects will attend to each other, while others will be ignored.

It is noteworthy that instance segmentation models do not always generate maps covering all image pixels. Pixels that have not been annotated with any instance labels are considered as background and are annotated with a corresponding label. In that way, we allow the network to focus on regions with no detected instances, which is especially useful in forgery cases where entire objects have been removed from the original image, i.e., inpainting. The attention regions defined by the object-guided attention mask for a sample case are displayed in Fig. 3.

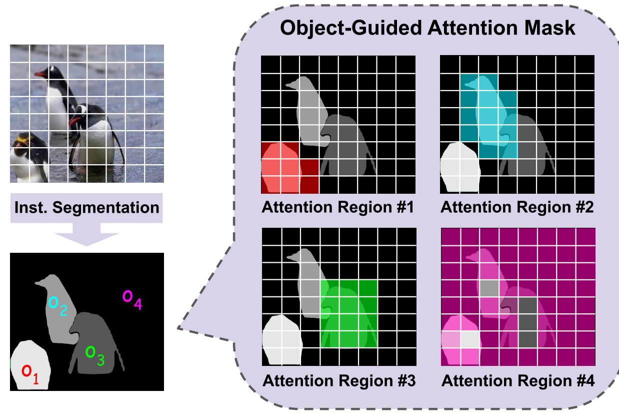


Figure 3. Object-Guided Attention Mask: Limits the attention of the transformer only between patches that depict the same objects. The four attention regions defined by the mask for an example image are depicted to the right. The background is considered as another object. For illustration purposes, the number of patches on both axes has been limited to eight.

Object-Guided Attention (OGA): We extend the self-attention mechanism [56] with our object-guided attention masks to guide tokens to attend only to tokens belonging to the same objects and ultimately derive object-level representations. Let an arbitrary token sequence $z \in \mathbb{R}^{L \times D}$ and its queries, keys and values $Q, K, V \in \mathbb{R}^{L \times D_h}$ be its projections for the self-attention. D_h denotes the dimensionality of the latent representation of the attention process. Then, the OGA calculation can be done as follows

$$Q = zW_q, \quad K = zW_k, \quad V = zW_v \tag{2}$$

$$OGA(z) = \sigma((QK^T + M)/\sqrt{D_h})V, \tag{3}$$

where $W_q, W_k, W_v \in \mathbb{R}^{D \times D_h}$ are the projection matrices, and σ denotes the softmax function. In that way, our masks force this process to attend only to token pairs with zero-value elements in the mask M.

In our implementation, we follow the multi-head version of self-attention [56]. To build our OGT, we use transformer blocks similar to ViT [14]. We first apply our OGA mechanism, followed by a Feed Forward network that acts on each token separately, consisting of an MLP with GELU activation [26]. Both operations are applied with a residual connection and a Layer Normalization [3] before their application. Finally, each forensic signal x_i^{sig} is processed by a dedicated stream that extracts patch representations that are processed by an OGT, consisting of B_1 transformer blocks, which outputs object-level representations $z_i^{sig} \in R^{L \times D}$ for the image patches.

RGB Stream: To encode and analyze the RGB image, we opt for using transformer-based backbone networks pretrained on large-scale image collections [25, 56] to capture relevant information in the RGB domain. Specifically, the x^{rgb} image is propagated through a pretrained vision backbone network [49], which composes our *RGB Stream*, to extract representations for the image regions. The output of this processing is denoted as $z^{rgb} \in R^{L\times D}$.

3.3. Token Fusion Module

Up to this point, the network has processed each of the N forensic signals and the RGB image using a separate stream, in a total of $N+1$ network streams. Practically, we have $N+1$ tokens for each image patch; hence, our goal is first to fuse all this information to a single representation for each patch and then capture the relation between the fused patch tokens. To this end, we propose the *Token Fusion Module (TFM)*, which comprises two main components: (i) the *Token Fusion Transformer (TFT)* for satisfying the former requirement, and (ii) the *Long-range Dependencies Transformer (LDT)*, for satisfying the later requirement. In that way, we transform the object-level features to image-level ones, which can be utilized for various downstream image forensics tasks.

Token Fusion Transformer (TFT): To shape the patch tokens generated from the RGB stream and the FSS into a form that our transformer can process, we stack all tokens together as follows:

$$z = [z_1^{sig}; z_2^{sig}; ...; z_N^{sig}; z^{rgb}] \quad \in \mathbb{R}^{(N+1)\times L\times D} \tag{4}$$

where z is the patch token tensor and $[\cdot;\cdot]$ denotes the stacking function on the outermost dimension. Hence, the $N+1$ tokens originating from the same image patch but from different streams are organized together. To fuse patch tokens derived from the same image patch, we follow the common practice in transformer literature [14, 56]: in particular, we employ a learnable *fusion token* $z^{ft} \in \mathbb{R}^{L\times D}$ repeated and stacked with the other patch tokens for each image patch. The goal is to refine the fusion tokens through the attention process of the TFT so as to incorporate relevant information from the $N+1$ tokens generated by the different streams. For the implementation of the TFT, we use B_2 transformer blocks [14]. Also, the TFT attention is applied to the outermost dimension of tensor z, which is considered the token sequence dimension. This can be formulated as follows:

$$[\bar{z}^{ft}; \bar{z}] = TFT([z^{ft}; z]), \tag{5}$$

where $\bar{z}^{ft} \in \mathbb{R}^{L\times D}$ are the *fused tokens*. After TFT, only the $\bar{z}^{ft}$ is considered for further processing, and the refined tensor $\bar{z}$ is discarded.

Long-range Dependencies Transformer (LDT): The fused tokens at the TFT output incorporate the forensic information coming out of all the $N+1$ streams. However, the information is aggregated so far only at the object level. Hence, we introduce the LDT component to take into account the forensic information at a global level and suppress wrongly captured inconsistencies while highlighting subtle but crucial traces. The LDT employs the transformer architecture [14] to refine the fused tokens $\bar{z}^{ft}$ into *forensic tokens* $z^{for} \in \mathbb{R}^{L\times D}$, by enabling the attention between all the L tokens. In that way, forensic tokens incorporate image-level information that derives from the relations between them, capturing long-range dependencies. For the implementation of LDT, a total of B_3 transformer blocks [14] are used.

3.4. Output Generation

To generate the outputs for the two target forensics tasks, we employ two specialized heads to further process the forensic tokens z^{for}. The first is the localization head, consisting of five transposed convolution layers, that generates the forgery localization output $\hat{y}^{loc}$. The second is the detection head, employing a four-block transformer [14] and a fully connected layer to predict the detection score $\hat{y}^{det}$.

3.5. Training process

Loss functions: To equally weight the error for the pristine and tampered regions, despite their expected different size on each sample, we employ the balanced Binary Cross Entropy (bBCE) loss $\mathcal{L}_{bBCE}$, for equalizing the contribution of both regions in the loss. We compute the bBCE by averaging the binary cross entropy loss computed separately for each of the two regions, i.e., pristine and tampered. Furthermore, to generate cleaner forgery localization masks, we also employ the Dice loss [24, 46] $\mathcal{L}_{dice}$ into the localization loss $\mathcal{L}_{loc}$. Regarding forgery detection loss $\mathcal{L}_{det}$, we employ the BCE loss. Thus, with a, b and c being weighting hyperparameters, the total training objective becomes

$$\mathcal{L} = \underbrace{(a\cdot\mathcal{L}_{bBCE} + b\cdot\mathcal{L}_{dice})}_{\mathcal{L}_{loc}} + c\cdot\mathcal{L}_{det}. \tag{6}$$

Stream Drop: On given training sets, some signals perform better than others. Thus, if we naively combine all streams, the network will learn to over-attend almost exclusively on the best-performing ones in the training set, which can significantly limit the generalization performance of the model. To mitigate this issue, inspired by the idea of DropPath [30] proposed for dropping residual paths, we introduce the *Stream Drop* (SD) mechanism for dropping entire streams of multi-stream networks during training to force the network to learn to capture useful information out of any input stream. To this end, each token in Eq. (4) is redefined based on the SD mechanism, which randomly drops a

	Approach	CASIAv1+		Columbia		Coverage		NIST16		OpenFor.		CocoGl.		DID		Overall	
		F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Feature Fusion	**SPAN** [28]	21.4	83.3	72.5	**98.1**	42.5	86.5	0.4	56.2	0.2	52.8	20.8	75.0	7.5	61.5	23.6	73.3
	IFOSN [61]	58.9	90.3	51.9	89.9	20.1	68.6	25.6	76.1	16.1	66.3	34.7	74.8	11.7	55.8	31.3	74.5
	PSCC-Net [40]	62.1	84.2	26.2	66.8	29.2	80.1	10.9	58.8	18.1	60.8	48.2	83.3	44.6	74.7	34.2	72.7
	MVSSNet++ [12]	58.5	81.9	63.1	86.9	46.0	85.9	24.5	78.3	20.3	69.6	44.5	81.5	19.1	56.0	39.4	77.1
	TruFor [24]	71.7	95.7	67.6	95.9	52.7	89.4	28.2	76.0	69.4	88.8	44.3	86.4	42.9	83.2	53.8	87.9
	CATNetv2 [37]	70.2	96.3	83.4	95.2	35.0	75.7	16.0	62.4	70.1	81.8	41.2	78.0	75.2	95.2	55.9	83.5
	OMG-Fuser$_F$ (Ours)	**84.5**	**97.2**	**86.1**	97.2	**63.1**	**91.4**	**34.2**	**79.1**	**72.2**	**92.1**	**53.9**	**88.3**	**77.0**	**95.4**	**67.3**	**91.5**
Score Fus.	**DST-Fusion** [20]	75.3	94.8	85.4	93.5	48.8	79.2	14.7	56.9	20.4	50.7	38.6	80.9	24.9	72.8	44.0	75.6
	AVG-Fusion	77.5	97.3	87.6	**98.9**	52.1	91.3	18.5	**83.1**	27.2	90.8	42.8	87.8	19.0	90.5	46.4	91.4
	OW-Fusion [7]	78.8	97.0	85.8	96.0	47.7	88.5	31.7	74.4	70.7	87.4	48.0	80.7	53.7	90.2	59.5	87.7
	OMG-Fuser$_S$ (Ours)	**85.1**	**98.1**	**92.9**	98.7	**70.1**	**96.0**	**37.5**	82.1	**74.1**	**94.1**	**56.2**	**89.4**	**76.6**	**96.1**	**70.4**	**93.5**

Table 1. Comparison on image forgery localization. Pixel-level F1 and AUC scores are presented for each algorithm and dataset. The best value per column is highlighted in bold, and the second best is underlined.

	Approach	CASIAv1+		Columbia		Coverage		NIST16		CocoGl.		Overall	
		F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Feature Fusion	**PSCC-Net** [40]	54.0	86.9	26.6	80.8	29.7	65.7	63.6	62.0	52.9	77.8	45.4	74.6
	TruFor [24]	80.1	91.7	98.3	99.6	56.2	77.0	28.7	62.7	48.6	75.2	62.4	81.2
	SPAN [28]	61.8	74.5	98.4	**99.9**	75.8	82.6	16.3	56.7	71.7	78.0	64.8	78.3
	IFOSN [61]	69.7	73.9	67.3	88.2	65.7	55.7	67.2	66.4	62.2	61.1	66.4	69.0
	MVSSNet++ [12]	72.2	85.5	78.1	98.3	66.1	71.3	65.8	58.9	67.6	67.9	70.0	76.4
	CATNetv2 [37]	86.1	94.4	83.1	95.2	64.0	67.9	69.4	75.0	64.8	66.6	73.5	79.8
	OMG-Fuser$_F$ (Ours)	**90.7**	**96.5**	**99.2**	**99.9**	**78.8**	**83.5**	**71.4**	**78.2**	**72.0**	**81.7**	**82.4**	**88.0**
Score Fus.	**DST-Fusion** [20]	80.3	93.83	92.3	97.9	69.0	78.4	44.4	75.1	66.0	78.5	70.4	84.7
	OW-Fusion [7]	85.3	93.3	80.1	97.7	66.1	72.4	69.8	75.5	69.3	72.2	74.1	82.2
	AVG-Fusion	86.1	93.0	94.5	99.7	73.3	80.7	69.0	76.7	69.2	78.2	78.4	85.7
	OMG-Fuser$_S$ (Ours)	**91.0**	**98.0**	**99.4**	**99.9**	**81.0**	**84.5**	**71.9**	**80.2**	**72.6**	**82.7**	**83.2**	**89.5**

Table 2. Comparison on image forgery detection. Image-level F1 and AUC scores are presented for each algorithm and dataset. The best value per column is highlighted in bold, and the second best is underlined.

stream with probability p_{drop}. SD can be formulated as

$$z_k = \begin{cases} z_k / p_{drop} & P(1 - p_{drop}) \\ 0 & P(p_{drop}) \end{cases} . \qquad (7)$$

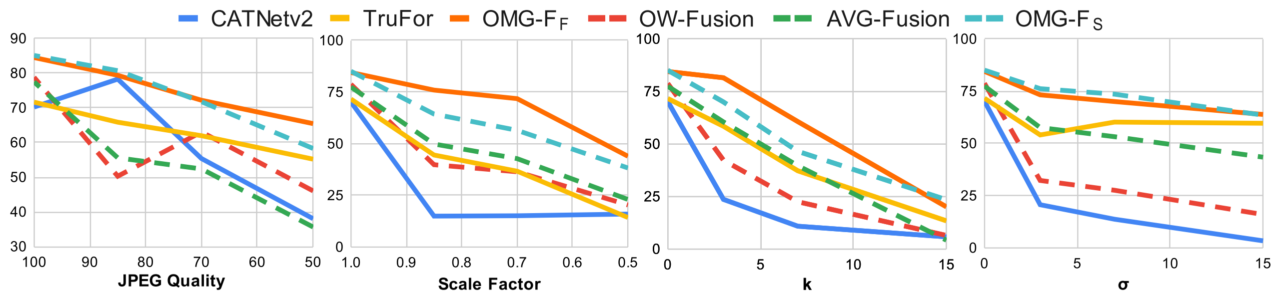
4. Experiments

4.1. Experimental setup

OMG-Fuser provides a network for combining both the score-level outputs of multiple image forensics algorithms as well as multiple low-level image forensic signals. Thus, we experiment with two different variants of our architecture: (i) one that incorporates five recent forgery localization algorithms [12, 24, 28, 37, 61] (OMG-Fuser$_S$), and (ii) one that directly fuses two recently proposed learnable forensic cues, namely the DCT stream [37] and the Noiseprint++[24] at the feature-level (OMG-Fuser$_F$). We train both variants on 40k forged samples and 32k authentic samples compiled from the datasets provided in [13, 36, 45], while we evaluate them on CASIAv1+ [12], Columbia [27], Coverage [59], NIST16 [23], OpenForensics [38], CocoGlide [24] and DID [60] datasets. We employ six recent competitor methods with publicly available implementations for the feature fusion category [12, 24, 28, 37, 40, 61]. Regarding score-level fusion, we reimplement two popular methods that support the fusion of multiple signals [7, 20], along with a baseline average fusion approach, using the same input signals as OMG-Fuser$_S$. We use the SAM [33] pretrained model for instance segmentation. DINOv2 [49] model is used as the vision backbone, which is fine-tuned on our OMG-Fuser$_F$ variant, but kept frozen on OMG-Fuser$_S$. Moreover, we employ the F1-score with a threshold of 0.5 as the main indicator of performance when deployed in the wild and the AUC as an auxiliary threshold-agnostic metric. More information regarding implementation details, datasets, and algorithms, along with localization results in terms of the F1 metric using the best threshold per image, is provided in the supplementary material.

4.2. Comparison with the state-of-the-art

Image forgery localization evaluation results are presented in Tab. 1. The reported F1 and AUC metrics are computed on the pixel-level. We see that our implementations outperform all state-of-the-art methods in score- and feature-level fusion. Moreover, in the case of score-level fusion, we show that previous approaches, either based on some statistical frameworks or learning approaches without semantic information, perform poorly on the recent deep learning-based forensic signals, even worse than averaging the sig-



(b) JPEG Compression (c) Resize (d) Gaussian Blur (e) Gaussian Noise

Figure 4. Robustness evaluation on common perturbations. The pixel-level F1 is reported. Straight lines denote the feature-level approaches, and dashed lines the score-level approaches. The top approaches of each category are shown for readability.

	Ablation Study	Loc. F1	Loc. AUC	Det. F1	Det. AUC
	OMG-Fuser$_S$	70.4	93.5	83.2	89.5
Signals	w/o SPAN	69.2	91.5	82.3	89.1
	w/o IFOSN	68.5	91.1	81.5	88.9
	w/o MVSSNet++	67.6	90.2	81.0	88.3
	w/o CATNetv2	63.6	89.7	78.1	87.6
	w/o TruFor	63.5	89.1	78.2	85.8
	w/o RGB	53.6	87.0	76.8	84.7
Compon.	w/o OGA	61.1	88.7	79.7	85.8
	w/o TFT	63.5	88.4	78.1	86.3
	w/o LDT	60.0	89.3	79.6	85.7
	w/o Stream Drop	66.5	90.5	81.1	87.3
	OMG-Fuser$_F$	67.3	91.5	82.4	88.0
Signals	w/o Noiseprint++	60.7	85.2	72.9	83.4
	w/o DCT	62.0	85.3	73.4	83.7
	w/o RGB	64.3	82.8	77.6	84.7

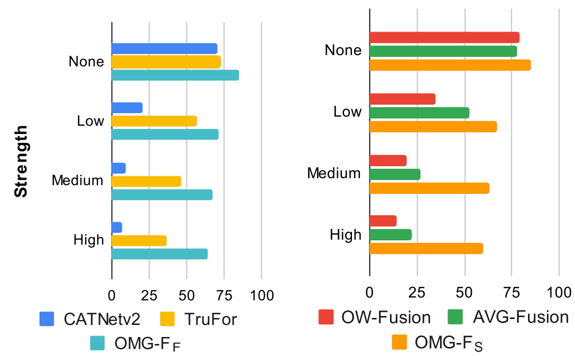
Table 3. Ablation Study. The average pixel-level F1 and AUC scores are reported across all evaluation datasets.

nals. This effectively shows that our network is capable of effectively fusing multiple forensic signals by exploiting the object-level information of the image, and at the same time, it is capable of handling low-level forensic traces. The high performance in feature-level fusion enables future works on new forensic signals to use our network to decrease the burden of developing purpose-built fusion architectures.

Image forgery detection evaluation results are presented in Tab. 2. Following [24], for methods that do not output an image-level forgery detection score, we compute it as the max of the corresponding forgery localization mask. Also, we omit datasets that contain only forged samples. Both OMG-Fuser implementations outperform all state-of-the-art methods by a significant margin.

4.3. Ablation Study

In order to better understand the contribution of each of the major components of our architecture, we performed several ablation studies in two directions. First, we evaluated the contribution of each of the fused forensic signals by training our network from scratch and removing the inputs



(a) Feature-level Fusion (b) Score-level Fusion

Figure 5. Robustness against neural filters for removing JPEG artifacts. The pixel-level F1 is reported.

of the network one at a time. Similarly, we proceeded by evaluating the contribution of each of the key components of our architecture by: replacing the OGA with the standard unmasked self-attention [56], replacing the TFT with an average pooling layer, removing the LDT and Stream Drop components. To better understand the impact of each component on signal fusion, we performed the later study on the OMG-Fuser$_S$, which fuses the output of several forensic algorithms. The results of the ablation study are presented in Tab. 3. Removing a single input signal affects the performance of the network adversely. Thus, the network is capable of learning to exploit even minor additional information introduced by new input streams. Moreover, we see that removing any key component of the architecture considerably impacts performance, highlighting that they are of crucial importance for effective signal fusion.

4.4. Robustness against Perturbations and Filters

To assess the robustness of our architecture against common online perturbations, we performed another set of experiments, where we applied different levels of JPEG compression, resizing, Gaussian noise, and Gaussian blur to the input images. We used the challenging CASIAv1+ dataset, as it contains already compressed samples with many different depicted objects. The results of this study for the task of image forgery localization are presented in Fig. 4, where

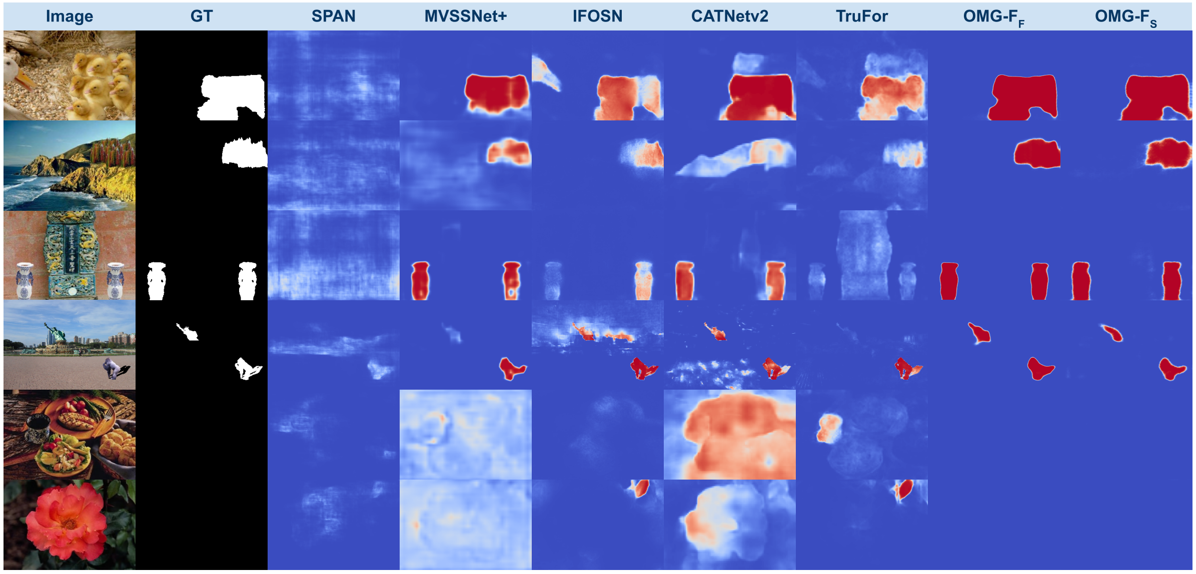


Figure 6. Qualitative evaluation results. From left to right, the image in question, the ground truth mask, the outputs of five recent methods, and the outputs of the two variants of our architecture are displayed.

the pixel-level F1 is reported. We see that the two variants of our architecture consistently outperform the state-of-the-art methods, both in feature-level and score-level fusion. Moreover, we evaluate our implementations against the recently added neural filter in Adobe Photoshop [1] for removing JPEG artifacts and present the results in Fig. 5. Our method demonstrates its robustness even in this challenging forgery attack, outperforming all competing approaches.

4.5. Stream Expansion

The design of the TFT allows the expansion of an already trained model with additional streams. In Tab. 4, we report the overall performance across all the evaluation datasets when expanding our architecture in three different ways. Starting from a variant of OMG-Fuser$_S$ trained without the TruFor stream, we expand it i) by training only the additional stream and keeping the rest of the network frozen, ii) by fine-tuning the network for 15% of the initial training epochs, and iii) by training the network from scratch. We see that in the first two cases, the network achieves a performance very close to third by requiring only a portion of the computational resources. This suggests that our model has learned to fuse various forensic signals effectively; therefore, it can be expanded with new signals without training from scratch.

4.6. Qualitative Evaluation

In Fig. 6, we present a qualitative evaluation of the two variants of our architecture on samples containing one, two, or no forged regions. In all the cases, both our variants lead to more robust results than all competing methods, clearly identifying the forged regions of the image and reducing false positives. The fusion process has a minimal effect on computation time, adding less than 100ms to the total process executed on commodity hardware. More details about timings are provided in the supplementary material.

Expansion	**Loc.**		**Det.**	
	F1	AUC	F1	AUC
4-stream model	63.5	89.1	78.2	85.8
stream-only train	68.8	92.1	82.4	88.6
full fine-tuning	69.0	92.7	82.7	89.1
training from scratch	70.4	93.5	83.2	89.5

Table 4. Evaluation of expanding the network with a new stream. The average scores are reported across all the evaluation datasets.

5. Conclusions

In this paper, we introduced a novel transformer-based network architecture for fusing an arbitrary number of image forensic signals based on the object-level information of the image. We demonstrated that both the effective fusion of signals and object-level information is essential for robust forensic analysis. Also, the modularity of our architecture was shown to be effective for score-level fusion of five recent image forgery localization algorithms and feature-level fusion of two recently proposed learnable forensic cues, outperforming all state-of-the-art methods on the tasks of image forgery detection and localization, while being robust to several traditional and novel forensic attacks. Finally, our network can facilitate future work as its expansion with new signals does not necessitate training from scratch.

Acknowledgments: This work was supported by the Horizon Europe vera.ai project (grant no.101070093), Junior Star GACR (grant no. GM 21-28830M), and EuroHPC for providing access to MeluXina supercomputer (grant no. EHPC-DEV-2023D03-008).

References

[1] Adobe. Jpeg artifacts removal filter, 2023. https://helpx.adobe.com/photoshop/using/quick-actions/jpeg-artifacts-removal.html, accessed 10th Oct. 2023. 8

[2] Ritu Agarwal and Om Prakash Verma. An efficient copy move forgery detection using deep learning feature extraction and matching algorithm. *Multimedia Tools and Applications*, 79(11-12):7355–7376, 2020. 2

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[4] Quentin Bammey, Rafael Grompone von Gioi, and Jean-Michel Morel. Image forgeries detection through mosaic analysis: the intermediate values algorithm. *Image Processing On Line*, 11:317–343, 2021. 2

[5] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300, 2019. 2

[6] Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Improved dct coefficient analysis for forgery localization in jpeg images. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2444–2447. IEEE, 2011. 2

[7] Polychronis Charitidis, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Ioannis Kompatsiaris. Operation-wise attention network for tampering localization fusion. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2021. 1, 3, 6

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 4

[9] Giovanni Chierchia, Sara Parrilli, Giovanni Poggi, Luisa Verdoliva, and Carlo Sansone. Prnu-based detection of small-size image forgeries. In *2011 17th International Conference on Digital Signal Processing (DSP)*, pages 1–6. IEEE, 2011. 2

[10] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019. 2

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

[12] 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*, 2022. 2, 6

[13] 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. IEEE, 2013. 6

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 5

[15] Mohamed A Elaskily, Heba A Elnemr, Ahmed Sedik, Mohamed M Dessouky, Ghada M El Banby, Osama A Elshakankiry, Ashraf AM Khalaf, Heba K Aslan, Osama S Faragallah, and Fathi E Abd El-Samie. A novel deep learning framework for copy-moveforgery detection in images. *Multimedia Tools and Applications*, 79:19167–19192, 2020. 2

[16] Hany Farid. Exposing digital forgeries from jpeg ghosts. *IEEE transactions on information forensics and security*, 4 (1):154–160, 2009. 2

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

[18] Pasquale Ferrara, Marco Fontani, Tiziano Bianchi, Alessia De Rosa, Alessandro Piva, and Mauro Barni. Unsupervised fusion for forgery localization exploiting background information. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015. 3

[19] Anselmo Ferreira, Siovani C Felipussi, Carlos Alfaro, Pablo Fonseca, John E Vargas-Munoz, Jefersson A Dos Santos, and Anderson Rocha. Behavior knowledge space-based fusion for copy–move forgery detection. *IEEE Transactions on Image Processing*, 25(10):4729–4742, 2016. 3

[20] Marco Fontani, Tiziano Bianchi, Alessia De Rosa, Alessandro Piva, and Mauro Barni. A framework for decision fusion in image forensics based on dempster–shafer theory of evidence. *IEEE transactions on Information Forensics and Security*, 8(4):593–607, 2013. 1, 3, 6

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

[22] Wenchao Gu, Shuang Bai, and Lingxing Kong. A review on 2d instance segmentation based on deep neural networks. *Image and Vision Computing*, 120:104401, 2022. 2

[23] 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. 6

[24] 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, 3, 5, 6, 7

[25] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 5

[26] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[27] Yu-Feng Hsu and Shih-Fu Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *2006 IEEE International Conference on Multimedia and Expo*, pages 549–552. IEEE, 2006. 6

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

[29] Yingying Hua, Ruixin Shi, Pengju Wang, and Shiming Ge. Learning patch-channel correspondence for interpretable face forgery detection. *IEEE Transactions on Image Processing*, 32:1668–1680, 2023. 2

[30] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016. 5

[31] Chryssanthi Iakovidou, Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. Content-aware detection of jpeg grid inconsistencies for intuitive image forensics. *Journal of Visual Communication and Image Representation*, 54:155–170, 2018. 2

[32] Gurpreet Kaur, Navdeep Singh, and Munish Kumar. Image forgery techniques: a review. *Artificial Intelligence Review*, 56(2):1577–1625, 2023. 1

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 6

[34] Paweł Korus and Jiwu Huang. Multi-scale fusion for improved localization of malicious tampering in digital images. *IEEE Transactions on Image Processing*, 25(3):1312–1326, 2016. 3

[35] Paweł Korus and Jiwu Huang. Multi-scale analysis strategies in prnu-based tampering localization. *IEEE Transactions on Information Forensics and Security*, 12(4):809–824, 2016. 3

[36] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 375–384, 2021. 6

[37] 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, 2, 6

[38] Trung-Nghia Le, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10117–10127, 2021. 6

[39] Zhouchen Lin, Junfeng He, Xiaoou Tang, and Chi-Keung Tang. Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis. *Pattern Recognition*, 42(11):2492–2501, 2009. 2

[40] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7505–7517, 2022. 6

[41] Yaqi Liu, Chao Xia, Xiaobin Zhu, and Shengwei Xu. Two-stage copy-move forgery detection with self deep matching and proposal superglue. *IEEE Transactions on Image Processing*, 31:541–555, 2021. 2

[42] Yaqi Liu, Binbin Lv, Xin Jin, Xiaoyu Chen, and Xiaokun Zhang. Tbformer: Two-branch transformer for image forgery localization. *IEEE Signal Processing Letters*, 2023. 2

[43] Jan Lukas, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2): 205–214, 2006. 2

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

[45] Gaël Mahfoudi, Badr Tajini, Florent Retraint, Frederic Morain-Nicolier, Jean Luc Dugelay, and PIC Marc. Defacto: image and face manipulation dataset. In *2019 27Th european signal processing conference (EUSIPCO)*, pages 1–5. IEEE, 2019. 6

[46] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 5

[47] Sophie J Nightingale, Kimberley A Wade, and Derrick G Watson. Can people identify original and manipulated photos of real-world scenes? *Cognitive research: principles and implications*, 2(1):1–21, 2017. 1

[48] Tina Nikoukhah, Jérémy Anger, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. Zero: a local jpeg grid origin detector based on the number of dct zeros and its applications in image forensics. *Image Processing On Line*, 11:396–433, 2021. 2

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5, 6

[50] Anh-Thu Phan-Ho and Florent Retraint. A comparative study of bayesian and dempster-shafer fusion on image forgery detection. *IEEE Access*, 10:99268–99281, 2022. 1, 2

[51] Alin C Popescu and Hany Farid. Exposing digital forgeries in color filter array interpolated images. *IEEE Transactions on Signal Processing*, 53(10):3948–3959, 2005. 2

[52] N Hema Rajini. Image forgery identification using convolution neural network. *International Journal of Recent Technology and Engineering*, 8(1):311–320, 2019. 2

[53] Victor Schetinger, Manuel M Oliveira, Roberto da Silva, and Tiago J Carvalho. Humans are easily fooled by digital images. *Computers & Graphics*, 68:142–151, 2017. 1

[54] Maria Siopi, Giorgos Kordopatis-Zilos, Polychronis Charitidis, Ioannis Kompatsiaris, and Symeon Papadopoulos. A multi-stream fusion network for image splicing localization. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part II*, pages 611–622. Springer, 2023. 3

[55] Yu Sun, Rongrong Ni, and Yao Zhao. Et: Edge-enhanced transformer for image splicing detection. *IEEE Signal Processing Letters*, 29:1232–1236, 2022. 2

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5, 7

[57] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. 1, 2

[58] 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*, 2022. 2, 3, 4

[59] 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. IEEE, 2016. 6

[60] Haiwei Wu and Jiantao Zhou. Iid-net: Image inpainting detection network via neural architecture search and attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1172–1185, 2021. 6

[61] 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. 1, 2, 6

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

[63] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34:30392–30400, 2021. 4

[64] Marcello Zanardelli, Fabrizio Guerrini, Riccardo Leonardi, and Nicola Adami. Image forgery detection: a survey of recent deep-learning approaches. *Multimedia Tools and Applications*, 82(12):17521–17566, 2023. 1, 2

[65] Ying Zhang, Jonathan Goh, Lei Lei Win, and Vrizlynn LL Thing. Image region forgery detection: A deep learning approach. *SG-CRC*, 2016:1–11, 2016. 2

[66] Yushu Zhang, Qing Tan, Shuren Qi, and Mingfu Xue. Prnu-based image forgery localization with deep multi-scale fusion. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–20, 2023. 2

[67] Lilei Zheng, Ying Zhang, and Vrizlynn LL Thing. A survey on image tampering and its detection in real-world photos. *Journal of Visual Communication and Image Representation*, 58:380–399, 2019. 1

[68] 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. 1, 2, 3

[69] Long Zhuo, Shunquan Tan, Bin Li, and Jiwu Huang. Self-adversarial training incorporating forgery attention for image forgery localization. *IEEE Transactions on Information Forensics and Security*, 17:819–834, 2022. 3