DOA-GAN: Dual-Order Attentive Generative Adversarial Network for Image Copy-move Forgery Detection and Localization

Ashraful Islam[1], Chengjiang Long[2]*, Arslan Basharat[2], and Anthony Hoogs[2]
[1]Rensselaer Polytechnic Institute, Troy, NY
[2]Kitware Inc., Clifton Park, NY
islama6@rpi.edu, {chengjiang.long, arslan.basharat, anthony.hoogs}@kitware.com

Abstract

Images can be manipulated for nefarious purposes to hide content or to duplicate certain objects through copy-move operations. Discovering a well-crafted copy-move forgery in images can be very challenging for both humans and machines; for example, an object on a uniform background can be replaced by an image patch of the same background. In this paper, we propose a Generative Adversarial Network with a dual-order attention model to detect and localize copy-move forgeries. In the generator, the first-order attention is designed to capture copy-move location information, and the second-order attention exploits more discriminative features for the patch co-occurrence. Both attention maps are extracted from the affinity matrix and are used to fuse location-aware and co-occurrence features for the final detection and localization branches of the network. The discriminator network is designed to further ensure more accurate localization results. To the best of our knowledge, we are the first to propose such a network architecture with the 1st-order attention mechanism from the affinity matrix. We have performed extensive experimental validation and our state-of-the-art results strongly demonstrate the efficacy of the proposed approach.

1. Introduction

The content of digital images can be easily manipulated or forged as there are many image editing tools like GIMP or Adobe Photoshop. Such manipulations can be done for nefarious purposes to either hide or duplicate an object or similar content in the original images. A copy-move image forgery refers to a type of image manipulation where a source region is copied to another location within the same image. As two real-world examples in Figure 1, copy-move image forgery could be used to add or hide some objects appearing a digital image, leading to a different interpretation. If such a manipulated image was part of a criminal investigation, without effective forensics tools the investigators could be misled. Therefore, it is crucial to develop a robust image forensic tool for copy-move detection and localization.

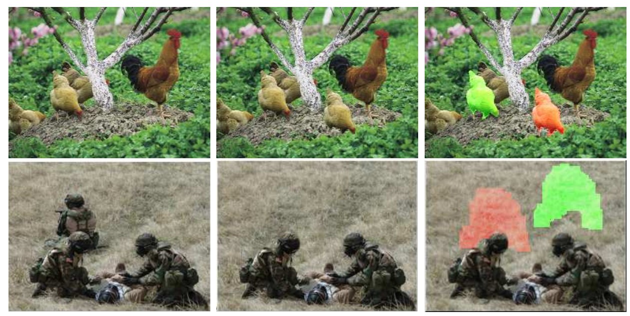

Figure 1: Two examples of copy-move forgery with object cloning (top) and object removal (bottom). From left to right are original, forged, and ground-truth images. Our goal is to automatically detect and localize the source (green) and the target (red) regions in forged images.

A number of copy-move detection approaches are already available including various traditional patch/block-based methods [8, 32, 17], keypoint-based methods [49, 33], irregular region-based methods [19, 36], and a few recent deep learning approaches [44, 22, 46]. Although some copy-move detection methods have been able to generate reasonable localization result, but the results of these approaches are still far from perfect on some of the more challenging scenarios. As shown in Figure 1, it is very challenging to distinguish copy-moves from incidental similarities, which occur frequently.

In this paper, we propose a dual-order attentive Generative Adversarial Network (DOA-GAN) for copy-move forgery detection and localization. As illustrated in Figure 2, the generator is an end-to-end unified framework based on a deep convolutional neural network. Given an input image, we calculate an affinity matrix based on the

*This work was supervised by Chengjiang Long when Ashraful Islam was a summer intern at Kitware, Inc.

extracted feature vectors at every pixel. We design a dual-order attention module to produce the 1st-order attention map A_1, which is able to explore the copy-move aware location information, and the 2nd-order attention map A_2 to capture more precise patch inter-dependency. The final feature representation is formulated with these two attention maps, and then fed into a detection branch to output a detection confidence score and a localization branch to produce a prediction mask in which the source region and target/forged region are distinguished. Meanwhile, the discriminator is designed to check whether the predicted mask is identical to ground-truth or not.

Intuitively, the dual-order attention module is designed to first highlight all similar regions in the image, regardless of whether or not they are manipulated; and then differentiate non-manipulated, similar regions from copy-move (source and target) regions. Typically, source and target regions in copy-move forgeries are more pixel-wise similar than incidentally similar regions, even after transformations such as rotation and scaling.

Our dual-order attention module is calculated based on the affinity matrix, which covers 2nd-order statistics of features and plays a critical roles for more discriminative representation [20, 9]. This motivates us to exploit the second-order co-occurrence attention map A_2 for the fine-grained distinctions necessary to distinguish copy-move forgeries from incidental object and texture similarities. Also, we observed that the high values in off-diagonal elements indicate high likelihood for copy-move spatial relations between the patches. This observation inspired us to explore the 1st-order attention map A_1 to focus on the copy-move region aware feature representation. In this paper, we refine and normalize the affinity matrix, taking the top-k values for each column and reshape them to form a 3D tensor with k channels. The tensor is then fed into simple convolutions to formulate our final 1st-order attention map A_1 which is able to give more attention to the source and target region. To the best of our knowledge, we are the first to extract such a 1st-order attention map from the affinity matrix.

We adopt the adversarial training process [13, 10, 43, 50] between the generator and the discriminator to generate a more accurate localization mask. As the number of epochs increases, both the generator and the discriminator improve their functionality so that the predicted mask iteratively becomes just like the ground-truth mask. Therefore, a sufficiently large number of epochs leads to convergence in training, and we use the learned parameters in the generator to output a detection confidence score and the predicted localization mask indicating source and target/forged regions.

To summarize, our contributions are three-fold. (1) We propose a dual-order attentive Generative Adversarial Network for image copy-move forgery detection and localization. (2) Our 1st-order attention module is able to extract the copy-move location aware attention map and the 2nd-order attention module explores pixel-to-pixel inter-dependence. These two attention maps provide more discriminative feature representations for copy-move detection and localization. (3) Extensive experiments strongly demonstrate that the proposed DOA-GAN clearly outperforms state-of-the-art approaches in terms of both detection and localization quality on multiple benchmark datasets.

2. Related work

Copy-move forgery detection and localization. A typical copy-move forgery detection approach [8] is composed of three stages: feature vector extraction, correspondence matching from the feature representation, and post-processing to reduce false alarms and improve detection rates. Patch/block-based methods include chroma features [3, 8], PCA feature [17], Zernike moments [39], blur moments [31], DCT [32]; keypoint-based methods such as SIFT [1, 7, 49], ORB [51], triangles [2], SURF [33, 40], and irregular region-based methods [19, 36]. Many traditional copy-move detection algorithms rely on strong assumptions about specific image characteristics like edge sharpness and local features. However, such assumptions are not always satisfied in the forged images, since other transformations like compression, resampling, or geometric transformations may hide traces of the manipulation.

Recently, deep neural networks (DNNs) have been applied to visual recognition [25, 16, 29, 28, 26, 27, 15], object detection and segmentation [14, 5, 30], as well as image and video forgery detection research [22, 14, 44, 29, 46, 24, 47, 4]. Especially, Wu *et al.* [46] introduced an end-to-end DNN solution to detect copy-move forged images with source/target localization with two separate branches. Unlike these DNN methods, our proposed DOA-GAN formulates both detection and localization as an end-to-end unified framework in the Generator network, where the 1st-order attention and the 2nd-order attention significantly improve the detection and localization performance.

Attentive Generative Adversarial Networks. Attention mechanisms have been successfully used in Generative Adversarial Networks [10, 48, 37]. Unlike the existing attentive GANs, the dual-order attention module in our DOA-GAN is dependent on the affinity matrix calculated from contextual feature representation.

3. Method

The framework of the proposed approach is illustrated in Figure 2. The generator is an end-to-end unified framework to conduct both copy-move manipulation detection and localization tasks. Given an input image I, we first apply the first four blocks of a VGG-19 network to extract hierarchical features and resize them to the same size to form a

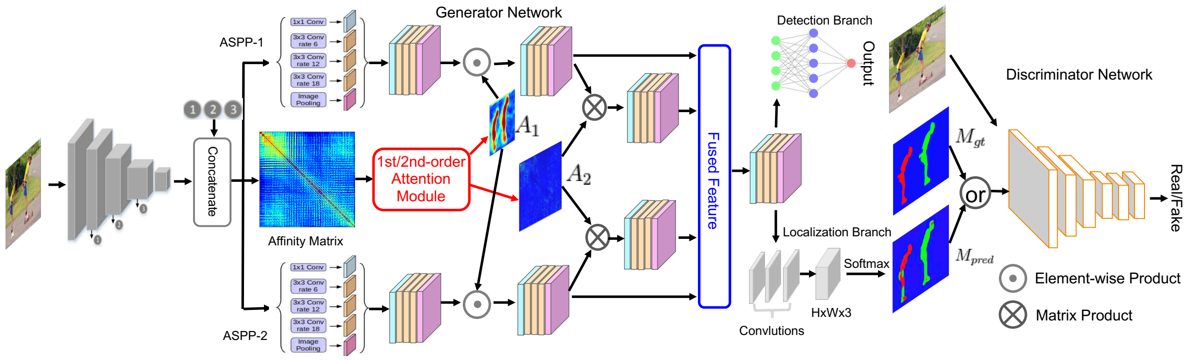

Figure 2: The overview of DOA-GAN. The generator is an end-to-end unified framework to conduct both detection and localization tasks. The discriminator is designed to check whether the predicted mask is identical to ground-truth or not.

concatenated feature F_{cat}. Then an affinity matrix is calculated, and the 1st-order attention map A_1 and the 2nd-order attention map A_2 are obtained via a dual-order attention module. Two atrous spatial pyramid pooling (ASPP) operations, *i.e.*, ASPP-1 and ASPP-2, with different parameters, are applied to extract contextual features F^1_{aspp} and F^2_{aspp}, which are multiplied element-wise with A_1 to get the possible copy-move regions attentive features F^1_{attn} and F^2_{attn}. A_2 is then used to obtain co-occurrence features F^1_{cooc} and F^2_{cooc}. Both region attentive features and co-occurrence features are fused for the detection branch to produce a detection output score and for the localization branch to generate a mask. The discriminator is designed to check whether the predicted mask is identical to the ground-truth or not. The alternative training between the generator and the discriminator is a key component of this approach and enables more accurate results.

3.1. Generator Network

Given an image $I \in \mathbb{R}^{H\times W\times 3}$, we extract feature representations of the image by feeding it to the first three blocks of a VGG-19 as feature extractor and then resize the three hierarchical features to the same size to get the concatenated feature $F_{cat} \in \mathbb{R}^{h\times w\times d}$. For time efficiency, we set $h = \frac{H}{8}$, $w = \frac{W}{8}$ in this paper. After feature extraction, to explore the correlation between different parts of the image, we calculate the affinity matrix

$$S = F'_{cat}F'^{T}_{cat}, \tag{1}$$

where $F'_{cat} \in \mathbb{R}^{hw\times d}$ is a flattened matrix representation of F_{cat} and represents $h \times w$ patches of the same size.

The Dual-Order Attention Module is designed as shown in Figure 3 to extract the copy-move aware region attention map A_1 and the co-occurrence attention map A_2. However, as we are calculating self-correlation of an image, S will have higher values along the diagonal, as the diagonal values indicate the correlation of a part of the image with itself. To resolve this issue, we define an operation G

$$G(i,j,i',j') = 1 - \exp(\frac{(i-i')^2 + (j-j')^2}{2\sigma^2}) \tag{2}$$

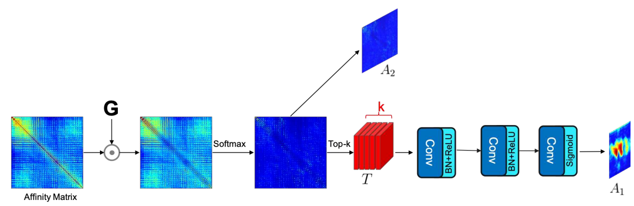

Figure 3: The dual-order attention module to obtain the copy-move region attention map A_1 and the co-occurrence attention map A_2.

and reshape it into $hw \times hw$. G reduces the correlation score between the same parts of the image using a Gaussian kernel. After that, we get the new affinity matrix $S' = S \odot G$, where $\odot$ denotes the element-wise product.

Leveraging the patch-matching strategy from [6], we calculate the likelihood that a patch in the i-th row matches with a patch in the j-th column in S' by

$$L^r(i,j) = \frac{\exp(\alpha S'[i,j])}{\sum_{j'=1}^{hw}\exp(\alpha S'[i,j'])}, \tag{3}$$

$$L^c(i,j) = \frac{\exp(\alpha S'[i,j])}{\sum_{i'=1}^{hw}\exp(\alpha S'[i',j])}, \tag{4}$$

$$L(i,j) = L^r(i,j)L^c(i,j), \tag{5}$$

where α is a trainable parameter that is initialized as 3. L is the final affinity matrix.

From $L \in \mathbb{R}^{hw\times hw}$, we extract the top-$k$ values for each row, and reshape into $T \in \mathbb{R}^{h\times w\times k}$. We feed T into an attention module. The attention module consists of three convolution blocks. The first two blocks contain convolution layers with 16 output channels and kernel size 3, followed by BatchNorm and ReLU. The final block contains two consecutive convolution layers with 16 output channels and kernel size 3, and 1 output channel and kernel size 1, respectively. We finally apply a sigmoid function to obtain the spatial copy-move aware attention map $A_1 \in \mathbb{R}^{h\times w}$. As illustrated in Figure 4, the copy-move region attention map is generated by suppressing the background non-manipulated regions while highlighting the regions most likely involved in a copy-move manipulation.

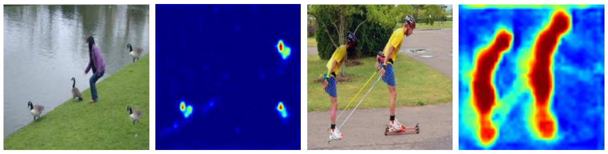

Figure 4: Visualization of A_1 on two copy-move forgery images.

To make full use of the patch-to-patch inter-dependence, we normalize the affinity matrix in Equation 5 to obtain co-occurrence attention map $A_2 \in \mathbb{R}^{hw \times hw}$,

$$A_2(i,j) = \frac{L(i,j)}{\sum_{j'=1}^{hw} L(i,j')}. \tag{6}$$

Atrous Spatial Pyramid Pooling (ASPP) Block is used to extract contextual feature from the extracted features F_{cat}. ASPP block is utilized in DeepLab V3 [5] to capture context at several ranges for image segmentation. We found through experiments that two ASPP blocks are useful to learn two different tasks, namely source and target detection. The first ASPP block has atrous rates 12, 24 and 36, and the second block has atrous rates 6, 12 and 24. After the ASPP modules, we obtain two feature representations $F^1_{aspp} \in \mathbb{R}^{h \times w \times d_s}$ and $F^2_{aspp} \in \mathbb{R}^{h \times w \times d_s}$.

Feature Fusion is designed to merge both copy-move region aware attentive features and co-occurrence features for the detection and localization tasks. We multiply F^1_{aspp} and F^2_{aspp} with the spatial copy-move region aware attention map A_1, and get

$$F^1_{attn} = F^1_{aspp} \odot A_1, \tag{7}$$

$$F^2_{attn} = F^2_{aspp} \odot A_1, \tag{8}$$

where $\odot$ is the element-wise product operation. We also obtain the co-occurrence features

$$F^1_{cooc} = A_2 \otimes F^1_{attn}, \tag{9}$$

$$F^2_{cooc} = A_2 \otimes F^2_{attn}, \tag{10}$$

where $\otimes$ is the matrix product operation. Such a treatment fully explores the inter-dependence between patches, and distant pixels are able to contribute to the feature response at a location based on similarity metrics.

The final feature representation is merged based on the above four features and attention map A_1,

$$F_{final} = \text{Merge}(F^1_{attn}, F^2_{attn}, F^1_{cooc}, F^2_{cooc}, A_1), \tag{11}$$

where Merge is merge operation. In principle, any kind of merge operation can be used, we used concatenation in this paper.

Detection Branch and Localization Branch. With the final representation F_{final}, we design two convolution layers followed by two fully connected layers as the detection branch to output a detection score. At the same time, F_{final} is fed into the localization branch, which consists of three convolution blocks, each followed by BatchNorm and ReLU, and a final convolution block of 3 channels to output the segmentation mask of pristine (background), source and target regions.

3.2. Discriminator Network

The structure of the discriminator is based on the PatchGAN discriminator [18]. Specifically, the discriminator is designed to predict whether each $N \times N$ patch in the image is real or fake. The discriminator is fully convolutional. It consists of five convolution blocks, each followed by BatchNorm and LeakyReLU, and a final convolution layer. The output channels of the consecutive convolution layers are 32, 64, 128, 256, 512, and 1, respectively, and the kernel size for all the convolution layers is 4×4. The stride of the convolution layers is 2 except the last one, which has a stride of 1. Therefore, as the input image is passed through each convolution block, the spatial dimension is decreased by a factor of two, and finally we get an output feature of size $\frac{H}{2^5} \times \frac{W}{2^5} \times 1$, where the spatial size of the input is $H \times W$. The input to the discriminator network is the concatenation of the image $I \in \mathbb{R}^{H \times W \times 3}$ and mask $M \in \mathbb{R}^{H \times W \times 3}$. The discriminator is trained to discern the ground-truth mask from the predicted mask, while the generator tries to fool the discriminator.

3.3. Loss Functions

The loss function is formulated with adversarial loss, cross-entropy loss, and detection loss as:

$$\mathcal{L} = \mathcal{L}_{\text{adv}} + \alpha \mathcal{L}_{\text{ce}} + \beta \mathcal{L}_{\text{det}}. \tag{12}$$

Adversarial Loss $\mathcal{L}_{\text{adv}}$ is defined as:

$$\mathcal{L}_{\text{adv}}(G, D) = E_{(I,M)}[\log(D(I,M)) + \log(1 - D(I, G(I)))], \tag{13}$$

where the discriminator D tries to maximize the objective, and the generator G tries to minimize it, *i.e.*,

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{adv}}(G, D). \tag{14}$$

Cross-Entropy Loss $\mathcal{L}_{\text{ce}}$ is expressed as:

$$\mathcal{L}_{\text{ce}} = \frac{1}{H \times W \times 3} \sum_{k=1}^{3} \sum_{i=1}^{H} \sum_{j=1}^{W} M(i,j,k) \log \widehat{M}(i,j,k), \tag{15}$$

where $\widehat{M} = G(I)$ is the predicted mask of the generator network, and M is the ground-truth mask.

Detection Loss $\mathcal{L}_{\text{det}}$ is the binary cross-entropy loss between the image-level detection score from the detection branch and ground truth label,

$$\mathcal{L}_{\text{det}} = y_{\text{im}} \log(\hat{y}_{\text{im}}) + (1 - y_{\text{im}}) \log(1 - \hat{y}_{\text{im}}), \tag{16}$$

Table 1: The copy-move forgery localization results on the USC-ISI CMFD dataset using pixel-level precision, recall, and F1 score metrics for 3 classes: P, S, and T referring to Pristine, Source and Target, respectively.

Methods	Precision			Recall			F1		
	P	S	T	P	S	T	P	S	T
BusterNet [46]	93.71	55.85	53.84	99.01	38.26	48.73	96.15	40.84	48.33
ManTra-Net [47]	93.50	8.66	48.53	**99.22**	2.28	28.43	96.08	2.97	30.58
U-Net [38]	91.66	32.67	47.16	97.16	19.06	40.90	94.88	23.09	44.15
NA-GAN	95.87	35.30	59.32	96.91	41.64	52.32	95.40	33.25	55.94
FOA-GAN	95.06	52.82	71.17	97.24	43.32	62.06	96.04	43.43	65.90
SOA-GAN	95.53	50.94	70.20	98.17	40.86	66.58	**97.80**	42.50	67.19
DOA-GAN w/o ASPP-1	96.71	61.04	70.94	98.84	43.13	66.69	97.67	45.04	67.23
DOA-GAN w/o ASPP-2	96.08	60.70	65.20	99.43	39.18	68.76	97.62	44.13	65.41
DOA-GAN w/o $\mathcal{L}_{\text{adv}}$	95.80	72.30	83.60	96.27	60.32	79.10	96.01	63.25	80.45
DOA-GAN w/o $\mathcal{L}_{\text{det}}$	**97.35**	75.58	83.96	97.98	**64.19**	80.31	97.51	65.21	81.08
DOA-GAN	96.99	**76.30**	**85.60**	98.87	63.57	**80.45**	97.69	**66.58**	**81.72**

where y_{im} is set to 1 if the image contains copy-move forgery, otherwise it is set to 0, and $\hat{y}_{\text{im}}$ is the output from the detection branch.

3.4. Implementation Details

The feature extraction module is based on the first three blocks of the VGG-19 network pretrained on the ImageNet dataset. The ASPP blocks are based on those used in DeepLabV3+ [5]. We used $k = 20$ for the top-k value in the 1st attention block.

We use two different learning rates for the generator and the discriminator networks, 0.001 and 0.0001, respectively, and the learning rate of the VGG-19 feature extractor is set to 0.0001. We decrease the learning rate by half when the training loss plateaus after 5 epochs. For training, we first optimize only the cross-entropy loss of the generator for 3 epochs, and then start optimizing all the losses. When the discriminator loss decreases to 0.3, we freeze the discriminator until the loss increases. This ensures that both the generator and the discriminator are learning at a similar pace, and the discriminator does not over-train.

4. Experimental Results

To study the effectiveness of the proposed DOA-GAN approach for copy-move forgery detection and localization, we conducted experiments on three benchmark datasets: the USC-ISI CMFD dataset [46], the CASIA CMFD dataset [46], and the CoMoFoD dataset [41].

The USC-ISI CMFD dataset has $80K$, $10K$, and $10K$ images for training, validation, and testing, respectively. The CASIA CMFD dataset contains $1,313$ forged images and their authentic counterparts (in total $2,626$ samples). The CoMoFoD dataset contains $5,000$ forged images, with 200 base images and 25 manipulation categories covering 5 manipulations and 5 post-processing methods.

For evaluation of detection and localization performance, we report image-level (for detection) and pixel-level (for localization) precision, recall, and F1 score metrics for 3 classes: Pristine (background), Source, and Target, by averaging the score of each image. The unit is %.

4.1. Experiments on the USC-ISI CMFD dataset.

We train DOA-GAN with $80,000$ copy-move forged images from USC-ISI dataset and $80,000$ pristine images, and evaluate on the $10,000$ testing forged images and $10,000$ pristine images. The pristine images are collected from COCO dataset [21]. We compare against BusterNet [46] as a baseline, because to the best of our knowledge, this is the only deep learning model that is able to distinguish between the copy-move source and target regions. To validate the effectiveness of the discriminator, we design several baselines, ManTra-Net [47], U-Net [38], DOA-GAN without any attention (denoted as NA-GAN), baselines using the 1st-order or 2nd-order attention only (denoted as FOA-GAN and SOA-GAN, respectively). We also created other baselines denoted as "DOA-GAN w/o $\mathcal{L}_{\text{adv}}$" (equivalent to DOA-CNN), and "DOA-GAN w/o $\mathcal{L}_{\text{det}}$", by removing the loss functions $\mathcal{L}_{\text{adv}}$, and $\mathcal{L}_{\text{det}}$ in Equation 12, respectively.

For pixel-level evaluation, we compute an average of precision, recall, and F1 score metrics for each image. As F1 score is ill-defined for pristine images, the testing images for pixel-level evaluation include only the forged images. For image-level evaluation, we use both forged images and non-forged images (total $20K$ images). We predict an image to be forged if the output score from detection branch is greater than 0.5, otherwise it is predicted to be non-forged. For BusterNet and DOA-GAN w/o $\mathcal{L}_{\text{det}}$, an image is considered forged if there are more than 200 pixels from the output mask predicted to be source or target regions. It is worth mentioning here that 200 pixels ($< 0.2\%$ of the total pixels in an input image of the size 320×320) is found to be a reasonable trade-off between the false negatives and false positives.

We have summarized the detection results in Table 2 and

Figure 5: Qualitative results on a sample from the USC-ISI CMFD dataset are shown. From left to right are input image; results of BusterNet [46], FOA-GAN, SOA-GAN, and DOA-GAN; and the ground truth mask, respectively. Note that the target region (as scaling transformation) is shown in red and the source region in green.

the localization results in Table 1. A few interesting observations: (1) DOA-GAN w/o $\mathcal{L}_{\text{adv}}$ performs better than BusterNet in terms of all the metrics, which clearly demonstrates promising performance for the generator in DOA-GAN; (2) DOA-GAN works better than DOA-GAN w/o $\mathcal{L}_{\text{adv}}$ overall in both detection and localization tasks, which demonstrates the efficacy of the discrimination ability from the discriminator in DOA-GAN, (3) The detection performance is worse in DOA-GAN w/o $\mathcal{L}_{\text{det}}$ than that in DOA-GAN, which demonstrates the efficacy of $\mathcal{L}_{\text{det}}$, (4) FOA-GAN and SOA-GAN perform worse than the DOA-GAN in all metrics except F1 score of pristine pixels, which suggests the 1st-order and the 2nd-order attentions are complementary to each other to improve the performance on the copy-move forgery detection and localization, and (5) U-Net and NA-GAN baselines perform much worse than DOA-GAN, SOA-GAN, and FOA-GAN, especially in localization of source mask, which demonstrates the efficacy of affinity computation. This indirectly verifies the effectiveness of our dual-order attention module. To further understand the advantage of the DOA-GAN approach, we also provide some visualization results in Figure 5. As we can see, our DOA-GAN is able to generate more accurate masks than BusterNet, our FOA-GAN, and our FOA-GAN.

Table 2: Detection results on the USC-ISI CMFD dataset.

Methods	Precision	Recall	F1
BusterNet [46]	89.26	80.14	84.45
ManTra-Net [47]	68.72	85.82	76.32
U-Net [38]	82.61	66.13	73.46
NA-GAN	80.19	85.64	82.82
FOA-GAN	94.13	94.54	94.33
SOA-GAN	95.50	92.30	93.87
DOA-GAN w/o ASPP-1	95.11	93.13	94.10
DOA-GAN w/o ASPP-2	92.97	91.75	92.35
DOA-GNN w/o $\mathcal{L}_{adv}$	95.45	93.09	94.25
DOA-GAN w/o $\mathcal{L}_{\text{det}}$	90.31	94.78	92.49
DOA-GAN	**96.83**	**96.14**	**96.48**

4.2. Experiments on the CASIA CMFD dataset.

Unlike the USC-ISI CMFD data, the CASIA CMFD dataset does not provide both ground-truth masks distinguishing source and target. It is more challenging because some uniform background is copied and pasted to the other background. To evaluate the proposed DOA-GAN on this dataset, we modified our network by replacing the final convolution layer of our network to a convolution layer of 1 channel output to get the mask of both copy and source parts as a single channel output. We train our model on the USC-ISI CMFD dataset and MS COCO dataset. For fair comparision, we do the same operation on BusterNet. In addition, we compare with four traditional copy-move forgery detection methods [1], *i.e.*, a block-based CMFD with Zernike moment features (denoted as "Block-ZM") [39], an adaptive segmentation based CMFD (denoted as "Adaptive-Seg") [36], a discrete cosine transform (DCT) coefficients based CMFD (denoted as "DCT-Match") [12], and a dense field-based CMFD (denoted as "DenseField") [8]. We evaluate the pixel-level performance by computing precision, recall, and F1 score metrics for each positive image where there is copy-move forgery, and report the final average. For image-level detection, we predict an image to contain forgery whenever there are more than 200 forged pixels in the output mask. We use both positive images and their authentic counterparts for image-level detection. All the images are resized to 320×320 before feeding into the models.

Table 3 shows performance comparisons with other baselines on CASIA CMFD dataset. As we can see, our proposed DOA-GAN performs the best in terms of all metrics except the precision in detection. This strongly demonstrates the promising advantages of our proposed method. Note that the result on BusterNet is different from the results reported in [46], as in the original BusterNet, the manipulation branch was trained on externel image manipulation datasets, whereas, for fair comparison, we train both our model and BusterNet only on the above-mentioned copy-move datasets.

Figure 6 provides a visualization result that shows the proposed DOA-GAN is able to detect more accurate masks than DenseField and BusterNet for the copy-move forgery manipulation.

4.3. Experiments on the CoMoFoD dataset.

We also evaluated the performance on the CoMoFoD dataset and report results in Table 4. Again, DOA-GAN

[1]Implementation available on https://github.com/MohsenZandi/Copy-Move_Forgery_Detection.

Figure 6: Visualization examples on the CASIA CMFD dataset. From left to right are the input image; results of Adaptive-Seg [36], DenseField [8], BusterNet [46], and our DOA-GAN; and the ground truth mask.

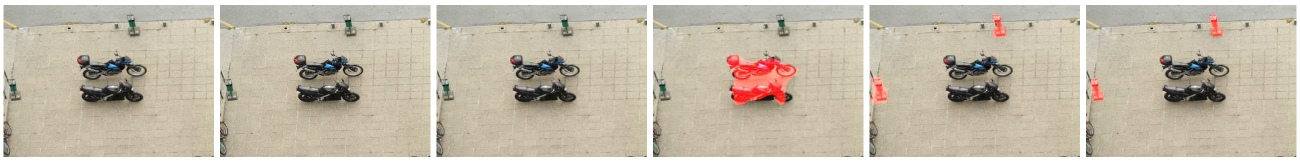

Figure 7: Visualization examples on the CoMoFoD dataset. From left to right are the input image; results of Adaptive-Seg [36], DenseField [8], BusterNet [46], and our DOA-GAN; and the ground truth mask.

Table 3: The performance on the CASIA CMFD dataset.

	Methods	Year	Precision	Recall	F1
Det	Block-ZM	2010	68.97	53.69	60.38
	DCT-Match	2012	63.74	46.31	53.46
	Adaptive-Seg	2015	93.07	25.59	40.14
	DenseFiled	2015	**99.51**	30.61	46.82
	BusterNet	2018	48.34	75.12	58.82
	DOA-GAN	2019	63.39	**77.00**	**69.53**
Loc	Block-ZM	2010	10.09	3.01	3.30
	DCT-Match	2012	8.80	1.90	2.40
	Adaptive-Seg	2015	23.17	5.14	7.42
	DenseField	2015	20.55	20.91	20.36
	BusterNet	2018	42.15	30.54	33.72
	DOA-GAN	2019	**54.70**	**39.67**	**41.44**

Table 4: The performance on the CoMoFoD dataset.

	Methods	Year	Precision	Recall	F1
Det	Block-ZM	2010	51.72	20.87	29.74
	DCT-Match	2012	50.48	29.77	37.46
	Adaptive-Seg	2015	65.66	43.37	52.24
	DenseField	2015	**80.34**	20.10	32.15
	BusterNet	2018	53.20	57.41	55.22
	DOA-GAN	2019	60.38	**65.98**	**63.05**
Loc	Block-ZM	2010	2.90	2.50	1.73
	DCT-Match	2012	3.53	3.41	2.03
	Adaptive-Seg	2015	23.02	13.27	13.46
	DenseField	2015	22.23	23.63	22.60
	BusterNet	2018	**51.25**	28.20	35.34
	DOA-GAN	2019	48.42	**37.84**	**36.92**

achieves the best performance except the precision in detection and localization. Note that different types of transformations are applied in this dataset to create copy-move manipulated images, *e.g.*, translation, rotation, scaling, combination, and distortion. Various post-processing methods, such as JPEG compression, blurring, noise adding, and color reduction, are also applied to all forged and original images.

Taking each post-processing method as a specific attack, we use this dataset to further analyze the effects of our proposed DOA-GAN under different attacks.

We provide a visualization example in Figure 7. Figure 9 shows the number of correctly detected images on CoMoFoD dataset under different types of attacks, where an image is correctly detected if its pixel-level F1 score is greater than 30%. Figure 8 shows F1 scores for all attacks. From these two figures, we can see that DOA-GAN is robust and consistently performs the best under all types of attacks.

4.4. Discussion

DOA-GAN is able to use the copy-move region attention to extract manipulation attentive features, as well as the co-occurrence feature with patch-to-patch interdependence taken into consideration. However, when the copy region is just extracted from the uniform background and pasted on the same background, it may fail. It also might fail when the scale has been changed significantly. We provide two failure cases in Figure 10. As we see, the backgrounds for the first example are uniform, and the scale of the copy-move regions are very small in the second example.

5. Extension to Other Manipulation Types

Note that DOA-GAN is based on an affinity matrix calculated on the same image. It is easy to extend it to an affinity matrix calculated from two different images, *i.e.*, donor image and probe image, and the corresponding manipulation types include image splicing and video copy-move.

For image splicing manipulation, we train DOA-GAN,

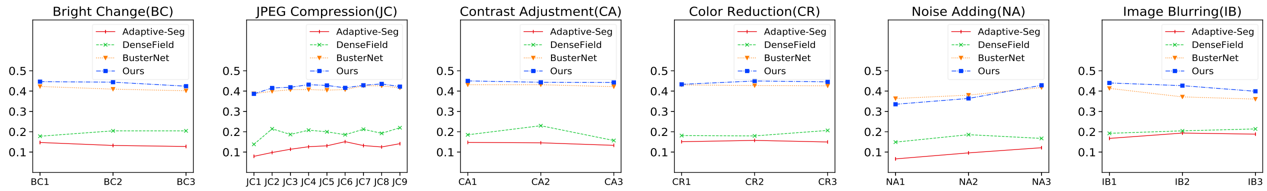

Figure 8: F1 scores on the CoMoFoD dataset under attacks.

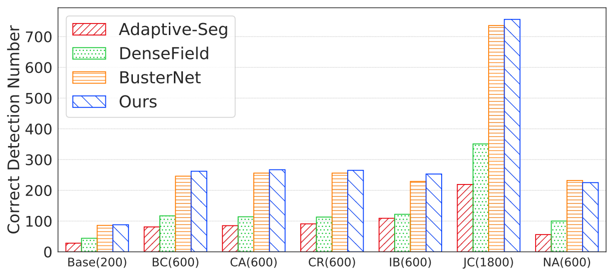

Figure 9: Number of correctly detected images on the CoMoFoD dataset under attacks. The total number of images for each attack is mentioned beside the name of the attack.

Figure 10: Examples of failure cases. From left to right - input image, our result, and ground truth.

and the two state-of-the-art approaches, DMVN [45] and DMAC [23], on the same synthetic image splicing dataset following the generation process in [23] and then evaluate on the generated dataset from MS-COCO consisting of 42,093 testing image pairs. The results in Table 5 demonstrate DOA-GAN's superiority in splicing localization.

For video copy-move manipulation, which can be considered as a inter-frame splicing between two consecutive frame sequences in a video. We evaluate our DOA-GAN's performance on the video copy-move datasets generated from video object segmentation datasets including DAVIS [34], SegTrackV2 [42] and Youtube-object [35], and summarize the results in Table 6. Again, the proposed DOA-GAN still works the best.

Note that due to page limit, we cannot provide sufficient technical details about extending DOA-GAN for image splicing and video copy-move. For more details of technical details, please refer to the supplementary.

Table 5: Performance comparison of image splicing localization on the generated dataset from MS-COCO.

Method	Source			Target		
	IoU	F1	MCC	IoU	F1	MCC
DMVN [45]	37.2	48.4	32.3	42.0	53.5	36.7
DMAC [23]	76.5	81.2	76.7	85.6	90.0	85.2
DOA-GAN	**86.4**	**91.0**	**86.2**	**92.4**	**95.4**	**91.8**

Table 6: Performance Comparison on the generated video CMFD dataset in terms of pixel-level F1 score and IoU. Here, S, and T, and A denote Source mask, Target mask, and source-target Agnostic mask, respectively.

Method	F1 Score			IoU		
	S	T	A	S	T	A
PatchMatch [11]	-	-	11.7	-	-	9.8
DMVN [45]	27.2	33.8	37.2	20.5	25.76	27.3
DMAC [23]	39.5	39.0	45.2	31.1	30.5	35.3
DOA-GAN	**62.9**	**62.3**	**65.0**	**50.7**	**49.6**	**53.3**

6. Conclusion and Future Work

In this paper, we propose a dual-order attentive Generative Adversarial Network (DOA-GAN) for copy-move forgery detection and localization. The dual-order attention module is designed in the generator to extract the manipulation location aware attention map and the underlying co-occurrence relations among patches. The discriminator is to further confirm the accuracy of the prediction masks. The proposed DOA-GAN has empirically shown to produce more accurate copy-move masks and better distinguish copy-move target regions from source regions, as compared to the previous state-of-the-art. Our future work includes extending the current work to identify image-level forgery in satellite images and solve other challenging vision tasks like co-saliency detection and localization.

Acknowledgement

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No.FA875016-C-0166. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

References

[1] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra. A SIFT-based forensic method for copy-move attack detection and transformation recovery. *TIFS*, 6(3):1099–1110, 2011.

[2] Edoardo Ardizzone, Alessandro Bruno, and Giuseppe Mazzola. Copy–move forgery detection by matching triangles of keypoints. *TIFS*, 10(10):2084–2094, 2015.

[3] Sevinc Bayram, Husrev Taha Sencar, and Nasir Memon. An efficient and robust method for detecting copy-move forgery. In *ICASSP*, 2009.

[4] Xiuli Bi, Yang Wei, Bin Xiao, and Weisheng Li. RRU-Net: The ringed residual U-Net for image splicing forgery detection. In *CVPRW*, 2019.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Jiaxin Cheng, Yue Wu, Wael Abd-Almageed, and Premkumar Natarajan. QATM: Quality-aware template matching for deep learning. In *CVPR*, 2019.

[7] Andrea Costanzo, Irene Amerini, Roberto Caldelli, and Mauro Barni. Forensic analysis of SIFT keypoint removal and injection. *TIFS*, 9(9):1450–1464, 2014.

[8] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Efficient dense-field copy–move forgery detection. *TIFS*, 10(11):2284–2297, 2015.

[9] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019.

[10] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *ICCV*, 2019.

[11] Luca D'Amiano, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. A patchmatch-based dense-field algorithm for video copy–move detection and localization. *TCSVT*, 29(3):669–682, 2018.

[12] A Jessica Fridrich, B David Soukal, and A Jan Lukáš. Detection of copy-move forgery in digital images. In *in Proceedings of Digital Forensic Research Workshop*. Citeseer, 2003.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[14] Xintong Han, Vlad Morariu, Peng IS Larry Davis, et al. Two-stream neural networks for tampered face detection. In *CVPRW*, 2017.

[15] Gang Hua, Chengjiang Long, Ming Yang, and Yan Gao. Collaborative active learning of a kernel machine ensemble for recognition. In *ICCV*, 2013.

[16] Gang Hua, Chengjiang Long, Ming Yang, and Yan Gao. Collaborative active visual recognition from crowds: A distributed ensemble approach. *T-PAMI*, 40(3):582–594, 2018.

[17] Deng-Yuan Huang, Ching-Ning Huang, Wu-Chih Hu, and Chih-Hung Chou. Robustness of copy-move forgery detection under high JPEG compression artifacts. *Multimedia Tools and Applications*, 76(1):1509–1530, 2017.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[19] Jian Li, Xiaolong Li, Bin Yang, and Xingming Sun. Segmentation-based image copy-move forgery detection scheme. *TIFS*, 10(3):507–518, 2014.

[20] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. Is second-order information helpful for large-scale visual recognition? In *ICCV*, 2017.

[21] 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 *ECCV*, 2014.

[22] Yaqi Liu, Qingxiao Guan, and Xianfeng Zhao. Copy-move forgery detection based on convolutional kernel network. *Multimedia Tools and Applications*, 77(14):18269–18293, 2018.

[23] Yaqi Liu, Xianfeng Zhao, Xiaobin Zhu, and Yun Cao. Adversarial learning for image forensics deep matching with atrous convolution. *arXiv preprint arXiv:1809.02791*, 2018.

[24] Chengjiang Long, Arslan Basharat, , and Anthony Hoogs. A coarse-to-fine deep convolutional neural network framework for frame duplication detection and localization in forged videos. In *CVPRW*, 2019.

[25] Chengjiang Long, Roddy Collins, Eran Swears, and Anthony Hoogs. Deep neural networks in fully connected crf for image labeling with social network metadata. In *WACV*, 2019.

[26] Chengjiang Long and Gang Hua. Multi-class multi-annotator active learning with robust gaussian process for visual recognition. In *ICCV*, 2015.

[27] Chengjiang Long, Gang Hua, and Ashish Kapoor. Active visual recognition with expertise estimation in crowdsourcing. In *ICCV*, 2013.

[28] Chengjiang Long, Gang Hua, and Ashish Kapoor. A joint gaussian process model for active visual recognition with expertise estimation in crowdsourcing. *IJCV*, 116(2):136–160, 2016.

[29] Chengjiang Long, Eric Smith, Arslan Basharat, and Anthony Hoogs. A c3d-based convolutional neural network for frame dropping detection in a single video shot. In *CVPRW*, 2017.

[30] Chengjiang Long, Xiaoyu Wang, Gang Hua, Ming Yang, and Yuanqing Lin. Accurate object detection with location relaxation and regionlets re-localization. In *ACCV*, 2014.

[31] Babak Mahdian and Stanislav Saic. Detection of copy–move forgery using a method based on blur moment invariants. *Forensic science international*, 171(2-3):180–189, 2007.

[32] Toqeer Mahmood, Tabassam Nawaz, Aun Irtaza, Rehan Ashraf, Mohsin Shah, and Muhammad Tariq Mahmood. Copy-move forgery detection technique for forensic analysis in digital images. *Mathematical Problems in Engineering*, 2016, 2016.

[33] VT Manu and Babu M Mehtre. Detection of copy-move forgery in images using segmentation and SURF. In *NeurIPS*. 2016.

[34] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

[35] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

[36] Chi-Man Pun, Xiao-Chen Yuan, and Xiu-Li Bi. Image forgery detection using adaptive oversegmentation and feature point matching. *TIFS*, 10(8):1705–1716, 2015.

[37] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *CVPR*, 2018.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[39] Seung-Jin Ryu, Min-Jeong Lee, and Heung-Kyu Lee. Detection of copy-rotate-move forgery using Zernike moments. In *International workshop on information hiding*, 2010.

[40] Ewerton Silva, Tiago Carvalho, Anselmo Ferreira, and Anderson Rocha. Going deeper into copy-move forgery detection: Exploring image telltales via multi-scale analysis and voting processes. *JVCIR*, 29:16–32, 2015.

[41] Dijana Tralic, Ivan Zupancic, Sonja Grgic, and Mislav Grgic. CoMoFoD–new database for copy-move forgery detection. In *ELMAR*, 2013.

[42] David Tsai, Matthew Flagg, and James M.Rehg. Motion coherent tracking with multi-label mrf optimization. *BMVC*, 2010.

[43] Jinjiang Wei, Chengjiang Long, Hua Zou, and Chunxia Xiao. Shadow inpainting and removal using generative adversarial networks with slice convolutions. *CGF*, 38(7):381–392, 2019.

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

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

[46] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. BusterNet: Detecting copy-move image forgery with source/target localization. In *ECCV*, 2018.

[47] Yue Wu, Wael Abd-Almageed, and Premkumar Natarajan. ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *CVPR*, 2019.

[48] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, pages 1316–1324, 2018.

[49] Bin Yang, Xingming Sun, Honglei Guo, Zhihua Xia, and Xianyi Chen. A copy-move forgery detection method based on CMFD-SIFT. *Multimedia Tools and Applications*, 77(1):837–855, 2018.

[50] Ling Zhang, Chengjiang Long, Xiaolong Zhang, and Chunxia Xiao. Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal. In *AAAI*, 2020.

[51] Ye Zhu, Xuanjing Shen, and Haipeng Chen. Copy-move forgery detection based on scaled ORB. *Multimedia Tools and Applications*, 75(6):3221–3233, 2016.