DoFNet: Depth of Field Difference Learning for Detecting Image Forgery

Yonghyun Jeong[1⋆], Jongwon Choi[2⋆], Doyeon Kim[1], Sehyeon Park[1], Minki Hong[1], Changhyun Park[1], Seungjai Min[1], and Youngjune Gwon[1]

[1] Samsung SDS, Seoul, Korea
{yhyun.jeong, dy31.kim, singing.park, mkidea.hong, arfken.park, seungjai.min, gyj.gwon}@samsung.com
[2] Dept. of Advanced Imaging, Chung-Ang University, Seoul, Korea
choijw@cau.ac.kr

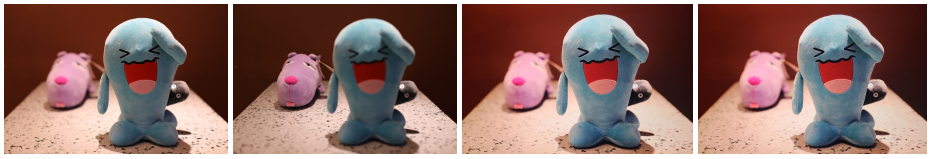
Abstract. Recently, online transactions have had an exponential growth and expanded to various cases, such as opening bank accounts and filing for insurance claims. Despite the effort of many companies requiring their own mobile applications to capture images for online transactions, it is difficult to restrict users from taking a picture of other's images displayed on a screen. To detect such cases, we propose a novel approach using paired images with different depth of field (DoF) for distinguishing the real images and the display images. Also, we introduce a new dataset containing 2,752 pairs of images capturing real and display objects on various types of displays, which is the largest real dataset employing DoF with multi-focus. Furthermore, we develop a new framework to concentrate on the difference of DoF in paired images, while avoiding learning individual display artifacts. Since DoF lies on the optical fundamentals, the framework can be widely utilized with any camera, and its performance shows at least 23% improvement compared to the conventional classification models.

1 Introduction

With rapid growth of vision technologies, online transactions have expanded its influence and even surpassed the proportion of offline transactions. Especially in the financial sector, various time-consuming and cumbersome offline procedures have been transformed into simple online procedures, including processing e-commerce payments, opening bank accounts, and filing for insurance claims.

Taking advantage of a contact-free environment of online transactions, scammers with malicious purposes use manipulated images for online frauds. Various schemes for image forgery are exploited, such as submitting other people's images as one's own for rental listing scams [1], or manipulating images using professional image editing tools, such as Adobe Photoshop, for a compensation fraud of railway delay [2]. As GAN-based synthesized images known as deepfakes have achieved highly realistic results, image forgery has become a potential threat in political, economic, and social aspects [3].

⋆ These authors contributed equally.



(a) Real image with center focus (b) Real image with background focus (c) Re-captured with center focus (d) Re-captured with background focus

Fig. 1: Real images and display images with different focus length

Fortunately, due to advancement in deep learning, the manipulating schemes for image forgery can be detected by the recent detection algorithms for deepfake images [4–7] and Photoshopped images [8] with superior performance.

When other people's images are re-captured and submitted as one's own, it is difficult to distinguish such cases by the current detection methods, since the re-captured images are also technically 'real' and not manipulated according to the current standards. To prevent such cases, many companies, especially in the insurance industry [9–11], provide mobile applications specifically developed for secure capturing and submission of images to file claims.

Unfortunately, some forgery methods are still available to fool detection, e.g., taking pictures of printed pictures or displayed objects on the screen. With the advancement of display panels with high resolutions, the display artifacts appear almost invisible when captured in images, which makes it challenging to distinguish between real images (Fig. 1(a)) and display images (Fig. 1(c)). From this point forward, we call the images taken of the real objects as *real images*, and the images taken of the displayed objects on the screen as *display images*. Regarding the current issues in online transactions, a new approach needs to be developed, not only to detect the cases of image forgery but also to prevent the initiation of frauds and scams.

In this paper, we propose a new approach to detect image forgery by analyzing the paired images of real and display, which are supposed to capture the difference in focal lengths as shown in Fig. 1. We begin our study with an intuition that the paired images of real objects contain variance in depth of field due to difference in focal lengths, while pictures of screen panels with displayed objects show consistency in depth. Based on the new dataset with over 2,700 paired images of real or displayed objects, we propose a novel framework for detection of the re-captured images of displayed objects.

The paired images in the dataset can be divided into two categories: the real images and display images. While the real objects are captured for real images, several monitors and a projector with displayed objects are captured for display images. With a detailed analysis of the dataset introduced in this paper, we observe that the appearance of the artifacts varies by the type of displays. Thus, if a classification model concentrates on the presence of the artifacts to distinguish the display images, the model cannot recognize the unknown artifacts outside of training settings. In such cases, failure in detection occurs.

To enhance the generality of the classification model across various types of displays, the proposed framework trains the model to concentrate on the difference in the variance of the depth of the paired images themselves, instead of the artifacts that may diversify based on the display types. With a detailed analysis of our new dataset, we validate the proposed framework through various ablation tests and confirm its superior performance on detecting the display images even with variance in display models across training and test phases.

This paper makes the following contributions:

- Our approach is the first study to detect pictures of display images by exploiting the difference between the paired images with a different focal length.
- We introduce a new dataset of 2,752 paired images, all of which are labeled as 'real' or 'display.'
- Based on the new network architecture processing the paired images simultaneously, we propose a new mechanism for the classification model to restrain from concentrating on the specific artifacts limited to the display types.
- With a detailed analysis of the new dataset and the superior performance of the proposed framework, we present the results of ablation analysis that validates the effectiveness of our approach.

Our dataset and source code are available online for public access. [1]

2 Related Work

In this section, we discuss the three lines of work most related to this paper: image forgery detection, depth of field, and multi-focus and focusing attention.

2.1 Image forgery detection

The rapid technological advancement of computer vision has made possible to produce high-quality forged images. Synthesized images are difficult to distinguish by naked eyes of human and have become almost impossible to detect without an in-depth analysis via trained AI models.

Currently, the most challenging forged images to detect is the GAN-based synthesized images known as deepfakes. With numerous generation models including ProGAN [12], BigGAN [13], CycleGAN [14], StarGAN 1,2 [15,16], StyleGAN 1,2 [17,18], deepfakes have improved to become highly realistic, and the object categories are expanded to include not only human faces but also animals like cats, dogs, and horses, and objects as automobiles, buildings, and paintings. Deepfake detection methods can be divided into two categories: natural characteristic based approach and synthesized artifact based approach. First, natural characteristic based approach focuses on the natural traits such as the details on head poses and movements [19], the absence of eye-blinking [20], the effect of biological signals as heart rate [21], and variance in lighting conditions

[1] https://github.com/SamsungSDS-Team9/DoFNet

and shadows [22–25]. Synthesized artifacts based approach focuses on observing the artifacts generated by GAN and can be categorized into pixel-based and frequency-based methods. The pixel-based method takes image pixels as an input of the classification network [4, 8, 26–32] while the frequency-based method converts the pixel domain (i.e., 2D data) into the frequency domain to take the frequency spectrum as an input of the classification network [5–7, 33–37].

Another image forensic method is to detect the image areas manipulated by Adobe Photoshop. Wang et al. [8] proposes a method for detecting facial warping using the liquifying tool on Adobe Photoshop and reconstructing the original image. Our approach is differentiated that we analyze and prevent image forgery from the initial stage of taking photos, instead of detecting forgeries after being manipulated already.

2.2 Depth of field

Popularized in photography and cinematography for directing the attention of the viewer, depth of Field (DoF) is an effect when objects within a certain range of distance appear clearly in focus and objects outside of the range, either closer or farther, appear blurry out of focus [38]. As in the work of Wu et al., DoF can be utilized for generating holographic imaging using autofocusing and phase recovery based on deep-learning [39]. Similarly, we also employ the difference of DoF in images to train our model for image forensics.

2.3 Multi-focus and Focusing Attention

Due to limited performance range of DoF of cameras, it can be difficult to take a picture with a clear focus on the entire image. In order to improve this issue, various frameworks have been studied regarding image focusing, including multi-focus and focusing attention mechanisms.

Multi-focus image fusion is fusing multiple images to produce an all-in-focus image. Guo et al. employed conditional generative adversarial network (cGAN) [40] for image-to-image fusion, which is known as FuseGAN [41]. Also, Zhang et al. introduced a large dataset containing realistic multi-focus images paired with their corresponding ground-truth images [42]. As Cheng et al. proposed, when focus is drifted to an unintentional region of the image, focusing attention mechanism can be employed to draw back the attention automatically using Focusing Attention Network (FAN) [43]. Different from the existing methods, we introduce a large dataset containing paired images intentionally focusing in the center and the background of the scene, respectively, and their corresponding display images captured in the same fashion.

To our knowledge, it is the first framework capable of analyzing image forensics from the moment when photos are taken. We achieved superior performance by employing a unique approach to develop our classification model based on the large dataset collected for this study.

3 Depth of Field Dataset

In this section, we introduce our new dataset containing paired images captured with different focal lengths to contain Depth of Field (DoF). DoF shows the variance in depth, especially when the objects located within a specific range of distance called midground appear in focus, while the other objects outside of the range called background appear blurry. The effect of DoF arises due to the physical properties of lenses. As the light passes through the lens of a camera or in our eye, the light source must keep a certain distance away from the lens to converge to a single point on the film or our retina. Based on DoF, we are enabled to not only predict the distances of various objects captured in images but also clearly express the concentrated areas or targeted objects.

3.1 Detecting Display Images using Paired Images with DoF

Conventionally, the target objects captured for e-commerce transactions are located at the center of the picture and the background is usually more distanced than the objects from the camera lens. Thus, in reality, each and every picture of real objects must contain DoF information. However, DoF captured in a single image is not a crucial factor in detecting image forgery, since both real and display images contain a certain level of DoF in the single image. To distinguish the real images from the display images, we utilize the paired images with variance in focal lengths. The one focusing in the center is called a *center image*, while the other focusing in the background is called *background image*. In the case of display images, the variance in depth would be relatively small between the center and the background images due to similar focal lengths from the camera lens to the display screen. On the contrary, the paired images of real objects contain a wide range of variance in depth between the center and the background regions.

It is beneficial in three aspects to exploit paired images with variance in depth. First, it does not require any additional sensor other than the camera lens itself, which indicates its simplicity in operation and scalability in any type of mobile device with embedded cameras. Second, it is not bound by any type of camera, since DoF is based on the physical properties of the camera lens. Finally, it is not based on the display artifacts, which indicates the generality of our model to accommodate any types of displays, including those unseen during the training phase. Therefore, by training the classification model to concentrate on the difference of DoF between the paired images, our model can distinguish the display images from the real images with superior performance. Unfortunately, the conventional classification model can be easily trained to focus on detecting the artifacts, which leads to limited generality of the classification model for the unknown artifacts in the test phase.

3.2 Data Collection

The DoF dataset is collected by five different models of mobile devices with the mobile application specifically developed to obtain the paired images with

Table 1: Comparison of various multi-focus datasets.

Dataset	Data acquisition method	Size(pair)	Resolution	Realistic	Ground Truth
Lytro [44]	Captured by light field camera	20	520×520	Yes	No
CNN [45]	Generated based on ImageNet dataset	1,000,000	16×16	No	Yes
BAnet [46]	Generated based on Matting dataset	2,268,000	16×16	No	Yes
FuseGAN [41]	Generated based on segmentation datasets	5,850	320×320	No	Yes
Real-MFF [42]	Captured by light field camera	800	625×433	Yes	Yes
Our dataset	Captured by cameras on mobile devices	2,752	720×1,280	Yes	Yes

Table 2: Classification performance with a single image and the paired images.

Single image classification							Paired image classification						
Train Display	Test Displays						Train Display	Test Displays					
	i-Mac		Samsung		Projector			i-Mac		Samsung		Projector	
	Acc.	A.P	Acc.	A.P.	Acc.	A.P.		Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
5K LCD	99.25	99.26	94.96	85.19	50.00	50.00	5K LCD	100.00	100.00	87.97	87.79	55.97	55.97
WQHD LED	50.38	51.12	100.00	100.00	76.52	76.52	WQHD LED	53.38	52.68	99.25	99.24	85.07	85.07
Projector	49.62	50.38	49.62	50.38	100.00	100.00	Projector	50.8	49.62	50.38	49.62	100.00	100.00

difference in depth. The application allows users to take two pictures at once as a package using the auto-focusing feature of the mobile camera. The paired images focus on the objects located at the center and the background, respectively. Among the various regions other than the center, the top of the center has the largest probability to be the background region. Thus, we set the focus setting to clearly capture the region at the top of the center by auto-focusing.

We construct the dataset into two categories: the paired images of real objects and the paired images of display objects. Since the display images should be similar to the real images, we first collect the paired images of real objects and then gather the paired images of display objects by re-capturing the real images displayed on the screens. To validate the algorithm with various target objects, we employ several object categories including shoes, cosmetics, music albums, DVD, household goods, and beverages. Furthermore, for the robust performance of the classification model in the various capturing environment, we have diversified the background settings, the distance to the target objects, and the capturing angle with the displays. To validate the robustness of the classification model with various unknown artifacts, we employ various models of displays for data collection, including a 5K LCD display of Apple iMac (Retina 5K, 27-inch, 2017), a WQHD LED display of Samsung monitor (LS27H850QFKXKR), and a display screen of NEC projector (NP-M311XG). In this way, we collect a large dataset composed of 2,752 pairs of images with $720 \times 1,280$ resolution for six different object categories. Each object category consists of four pairs of images: a pair of real images, and three pairs of display images obtained from a 5K LCD monitor, a WQHD LED monitor, and a projector screen, respectively. In addition, we compare the DoF dataset with other datasets containing paired images with variance in focal lengths. As indicated in Table 1, our dataset is the largest in size with the highest resolution of paired images among the realistic datasets, which demonstrates the scalability of DoF dataset in various tasks.

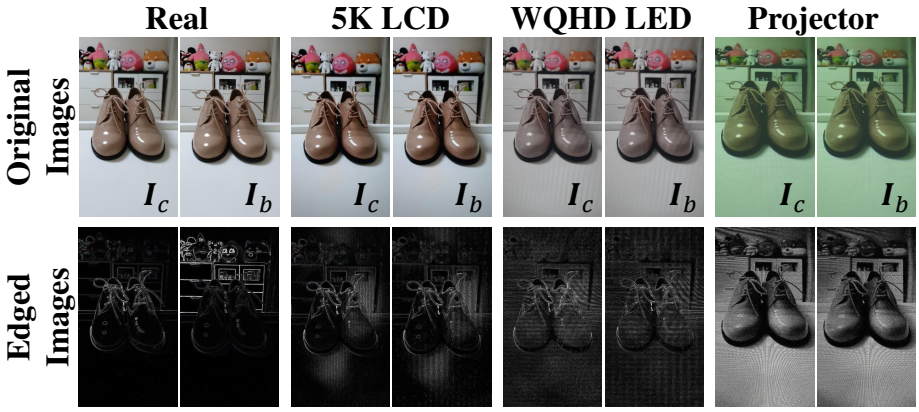


Fig. 2: Various display artifacts in DoF dataset

3.3 Analysis of Display Artifacts

As illustrated in Figure 2, some artifacts can be easily discovered in certain parts of the display images. The shapes of the artifacts are various according to the types of displays and the capturing angle with the display panel, so it is impossible to consider all kinds of artifacts in the training phase. Furthermore, it is challenging to train the model to consider every single new artifact whenever a brand-new display is launched.

To show the limited generality of the artifact-based classification, we train a neural network of ResNet-18 [47]. We utilize the center image of the paired images as the input of ResNet-18, which is a binary classification model that determines whether the input is a real image or a display image. This classification model with the single image is noted by *SingleNet-18*, and its detailed settings are given in Section 5.1.

As shown in the left side of Table 2, *SingleNet-18* has shown a great performance when the same displaying device is considered in both of the training and test phases, while the performance dramatically drops to almost 50% when new types of displays outside of training phase are employed in the testing phase. This result validates that artifact-based training should be avoided for generalizing the classification model across various displays.

A similar situation happens even when paired images are used as the input of the neural network. We extend the ResNet-18 to accept paired images as inputs. For the input of the paired images, the channel size of the input image of ResNet-18 has been expanded to 6 channels, and the input is obtained by concatenating the paired images in the channel dimension. This classification model with the paired images is noted by *DualNet-18*, and its settings and hyperparameters are all equivalent to *SingleNet-18*.

The results of *DualNet-18* are given in the right side of Table 2, which are similar to the results of *SingleNet-18*. Even when the paired images are given for the classification model, the wrong direction, cannot be corrected because the

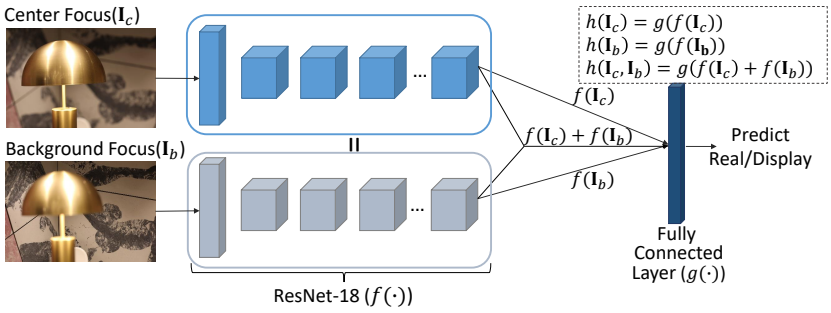


Fig. 3: Overall architecture of the proposed framework

artifacts are much easier to be trained than the variance of depth in the paired images. Thus, when we just train the classification model with the conventional training methods, the classification model drives to focus on the artifacts given in the training phase, which reduces the generality of the model dramatically.

4 DoF-based Detection of Display Images

We propose a new classification model and its training mechanism to concentrate on the variance in depth rather than the display artifacts. When the paired images are given for training, we denote the image with centered focus and the image focused on the background by $\mathbf{I}_c$ and $\mathbf{I}_b$, respectively.

4.1 Network Architecture

The proposed classification model contains two feature extractors and one fully-connected layer, as shown in Fig. 3. Each of the feature extractors is fed by $\mathbf{I}_c$ and $\mathbf{I}_b$, respectively. To let the feature extractors focus on the difference of DoF, the two feature extractors always share their weight parameters. Thus, we use the corresponding notation of $f(\mathbf{I})$ for the two feature generator, which means that the feature vector is the output for the given image of $\mathbf{I}$.

Then, the fully-connected layer determines the classes of the two feature vectors (i.e. $f(\mathbf{I}_c)$ and $f(\mathbf{I}_b)$) given from the feature extractors. By using $f(\mathbf{I}_c)$ and $f(\mathbf{I}_b)$, we consider the three combinations of the features. The first one is the dual combination where the two features are summed before the estimation of the fully-connected layer as $f(\mathbf{I}_c)+f(\mathbf{I}_b)$. The second and third combinations are the single ones where the individual features of $f(\mathbf{I}_c)$ and $f(\mathbf{I}_b)$ are fed into the fully-connected layer. When we denote the operation of the fully-connected layer by the function of $g(\bullet)$, the proposed network gives three outputs for one pair of images as: $g(f(\mathbf{I}_c)+f(\mathbf{I}_b))$, $g(f(\mathbf{I}_c))$, and $g(f(\mathbf{I}_b))$. For the simple representation, we denote the three outputs by $h(\mathbf{I_c}, \mathbf{I}_b)$, $h(\mathbf{I}_c)$, and $h(\mathbf{I}_b)$, respectively. By summing the two features before the fully-connected layer, we can obtain two advantages: first, since the two different features share the weight parameters

ofthe fully-connected layer, it results in a more balanced model considering the paired images simultaneously; second, we can avoid overfitting by reducing the size of the fully-connected classifier.

To reduce the computational load and the necessary resources, for the feature extractors, we select the ResNet-18 [47] that is the smallest model among the various ResNet models. Since the network architecture is different from the model with the single image, rather than using the pre-trained network, we initialize the weight parameters according to the He initialization [47].

4.2 Artifact-free Training Method

The objective of the proposed classification model is to distinguish the display images from the real images by using the difference of DoF in the paired images while ignoring the effect of artifacts. To consider the objective, we build the training loss as follows:

$$\mathcal{L}(\mathbf{I}_c, \mathbf{I}_b) = (1-\lambda)\mathcal{L}_{dual}(\mathbf{I}_c, \mathbf{I}_b) + \lambda\mathcal{L}_{dof}(\mathbf{I}_c, \mathbf{I}_b) \tag{1}$$

where λ is the scaling factor to control the effect of the two loss terms: $\mathcal{L}_{dual}$ and $\mathcal{L}_{dof}$. We call $\mathcal{L}_{dual}$ and $\mathcal{L}_{dof}$ by the dual classification loss and the DoF loss, respectively. $\mathcal{L}_{dual}$ works as the conventional classification loss to let the neural network classify the given pair of images well. In contrast, $\mathcal{L}_{dof}$ prevents the neural network from being trained by considering the display artifacts. The detailed role and the derivation of $\mathcal{L}_{dual}$ and $\mathcal{L}_{dof}$ are given in the following.

Dual Classification Loss. The dual classification loss is the fundamental loss to classify the real and display images according to their own labels. Thus, the dual classification loss is derived as:

$$\mathcal{L}_{dual}(\mathbf{I}_c^{(i)}, \mathbf{I}_b^{(i)}) = \texttt{CE}\left(h(\mathbf{I}_c^{(i)}, \mathbf{I}_b^{(i)}), l^{(i)}\right), \tag{2}$$

where $\mathbf{I}_c^{(i)}$ and $\mathbf{I}_b^{(i)}$ are respectively the center and the background images of the i-th pair of images, $l^{(i)} \in \{0, 1\}$ is their ground-truth label, and the $\texttt{CE}(y, l)$ means the softmax cross-entropy loss letting y go to the one-hot vector of l. In $l^{(i)} \in \{0, 1\}$, 0 represents the label for the real images, while 1 means the label for the display images. Thus, the dual classification loss drives the neural network to predict the correct labels of the given pair of images.

DoF Loss. For the neural network to ignore the effect of the artifacts, the DoF loss utilizes the two outputs from the single inputs, which are $h(\mathbf{I}_c)$ and $h(\mathbf{I}_b)$. Before deriving the single DoF loss, we first describe the difference between the center and background images. When we consider only the center images to detect the display images, the classification model focuses on the display artifacts, since the two images cannot be distinguished from each other without the artifacts. On the contrary, employing the background images of real paired

images enables the model to distinguish the display images without the artifacts, since the centered regions of the background images become blurry, in contrast to the display images. Thus, the center and the background images play different roles in the DoF loss, which results in two separated loss terms as follows:

$$\mathcal{L}_{dof}(\mathbf{I}_c, \mathbf{I}_b) = \frac{1}{2}\mathcal{L}_{dof-c}(\mathbf{I}_c) + \frac{1}{2}\mathcal{L}_{dof-b}(\mathbf{I}_b), \tag{3}$$

where $\mathcal{L}_{dof-c}$ is the DoF center loss that considers the center images only and $\mathcal{L}_{dof-b}$ indicates the DoF background loss utilizing only the background images.

Since the classification based on the center images lead the neural network to focus on the artifacts, we adversarially derive the DoF center loss as follows:

$$\mathcal{L}_{dof-c}(\mathbf{I}_c^{(i)}) = \texttt{CE}\left(h(\mathbf{I}_c^{(i)}), 1\right). \tag{4}$$

According to $\mathcal{L}_{dof-c}$, when only a single center image is given to the network, both of the real and display images are predicted by the same class of 1. Thus, the classification concentrating on the artifacts can be suppressed by the DoF center loss, since the loss drives the neural network to ambiguously detect the display images with only the center images.

In contrast to the center images, the background images are essential for training the difference of DoF. Thus, we derive the DoF background loss by the conventional classification loss as follows:

$$\mathcal{L}_{dof-b}(\mathbf{I}_b^{(i)}) = \texttt{CE}\left(h(\mathbf{I}_b^{(i)}), l^{(i)}\right). \tag{5}$$

After summing up the entire losses of Eq. 6 for the paired images sampled in the iterative mini-batch, we optimize the neural network by the momentum stochastic gradient descent algorithm [48]. Although the two distinct losses are summed up together, we do not consider any step-by-step training scheme for stable training.

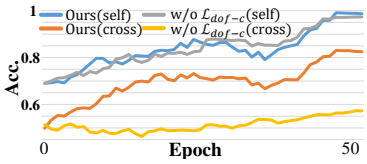


Fig. 4: Analysis for Loss

L_{dual} and L_{dof-c} contradict each other to adversarially train the backbone. To show the effectiveness of the adversarial training, we estimate the test accuracy at every epoch as illustrated in Fig. 4. While the proposed framework improves the cross-display and self-display accuracies, the cross-display accuracy dramatically declines without L_{dof-c}.

4.3 Implementation Details

Before the training phase, all the input images are pre-processed by resizing and random cropping operations. In resizing operation, the original image of 720×1280 pixels is resized into 256×256 pixels, and random cropping operation crops the image into 253×253 pixels from the resized image. Resizing operation

is applied to increase the computational efficiency of the neural network, while random cropping operation is necessary to cover the slight movement that can happen essentially during capturing the consecutive paired images. In the test phase, we consider $h(\mathbf{I}_c, \mathbf{I}_b)$ as the prediction of the given pair of images.

The proposed framework is optimized by the Stochastic Gradient Descent method (SGD) with the batch size of 16, executing 50 epochs, and the learning rate begins with 0.1 and later adjusts to 0.01 after 40 epochs. For the stable updates, we utilized the momentom SGD with the momentum of 0.9 and the decay weights of 5×10^{-4}. The λ, that is a scaling factor in Eq. 6, is set to 0.6.

5 Experimental Result

5.1 Setup of Experiments

Dataset Setup and Measurement. Based on the object categories, we split the dataset as follows: 80% as a training set, 10% as a validation set, and the last 10% as a test set. The first experiment trains the model with all of the real images and the display images of a single type of display in the training set. Then, the trained model is evaluated with the display images of two remaining types of displays in the test set. In the second experiment, the display images of two types of displays are utilized in the training phase, while the images of the remaining type of display are used for evaluation in the test phase. To effectively express the performance of our approach, we employ the measurements commonly used in deepfake detection: accuracy (Acc.) and average precision (A.P.) [4–6,8,19–21,30–32,35,37]. For experiments, we use GPUs of RTX Titan.

Table 3: Training of single display

Train Displays	Models	WQHD LED		Projector	
		Acc.	A.P.	Acc.	A.P.
5K LCD	SingleNet-18	84.96	85.19	50.00	50.00
	SingleNet-50	89.47	88.23	47.72	50.00
	DualNet-18	87.97	87.79	55.97	55.97
	DualNet-50	88.72	86.87	47.73	50.00
	Ours	**92.48**	**90.55**	**84.85**	**82.84**
Train Displays	Models	5K LCD		Projector	
		Acc.	A.P.	Acc.	A.P.
WQHD LED	SingleNet-18	50.38	51.12	76.52	76.52
	SingleNet-50	50.38	51.17	87.12	87.12
	DualNet-18	53.38	52.68	85.07	85.07
	DualNet-50	53.38	54.08	90.91	90.91
	Ours	**90.98**	**90.38**	**98.48**	**98.53**
Train Displays	Model	5K LCD		WQHD LED	
		Acc.	A.P.	Acc.	A.P.
Projector	SingleNet-18	49.62	50.38	49.62	50.38
	SingleNet-50	49.62	50.38	61.65	50.38
	DualNet-18	50.38	49.62	50.38	49.62
	DualNet-50	50.38	51.12	58.65	59.26
	Ours	**62.41**	**62.29**	**81.95**	**80.95**

Comparison of Models. For comparison of performance of our model with others, we designed the two simple networks described in Section 3.3. *SingleNet-18* exploits only a single image in a classifier, assuming the case of an image forgery by capturing other people's image as one's own. Similar to the backbone of our model, the classification model employs ResNet-18 as the CNN model for distinguishing the real and display images. Moreover, to assess performance with various depth of networks, ResNet-50 is also employed as *SingleNet-50*.

DualNet-18 exploits the paired images with variance in depth; by concatenating the two RGB images, the network considers six channels in total as an input to distinguish between the real images and the display images. We also extend *DualNet-18* by employing ResNet-50, is named as *DualNet-50*. The comparison models including *SingleNet-18*, *SingleNet-50*, *DualNet-18*, and *DualNet-50* are trained by ADAM optimizer [49], with the learning rate of 10^{-4} and the batch size of 16, executing the same number of epochs with the proposed framework. We utilize the ADAM optimizer for the comparison models since the SGD fails to train the neural network with the setting.

5.2 Experiment with Training of Single Display

In the DoF dataset, three types of displays are employed to obtain the display images. To show the robust performance of the framework on the unknown displays, we utilize only the display images captured on a single type of display in the training phase, while the remaining types of displays are employed during the test phase. Thus, we can perform the three experiments respectively utilizing the 5K LCD monitor, the WQHD LED, and the projector in the training phase.

Table 4: Training of multiple displays

Train display	Model	Acc.	A.P.
WQHD LED & Projector	SingleNet-18	54.89	54.20
	SingleNet-50	54.89	55.56
	DualNet-18	55.64	54.97
	DualNet-50	57.14	57.78
	Ours	**81.95**	**77.34**
5K LCD & Projector	SingleNet-18	82.71	82.44
	SingleNet-50	90.23	88.05
	DualNet-18	92.48	92.59
	DualNet-50	88.72	86.87
	Ours	**94.74**	**93.93**
5K LCD & WQHD LED	SingleNet-18	67.91	67.91
	SingleNet-50	66.67	65.97
	DualNet-18	61.19	61.19
	DualNet-50	60.61	58.40
	Ours	**93.28**	**92.56**

As shown in Table 3, the proposed framework has achieved the state-of-the-art performance in all cases. Even though *SingleNet-50* and *DualNet-50* utilize the deeper models than *SingleNet-18* and *DualNet-18*, the performance does not improve at all, which represents the depth of the neural network is not essential to detect DoF in paired images. As the proposed framework, *DualNet-18* and *DualNet-50* also utilize the paired images, their performance declines by 37.12% at most compared to our algorithm. From the result, we can confirm that the proposed framework concentrates on the variance in DoF in the given paired images, while avoiding the supervision affected by the display artifacts. Thus, our algorithm can be trained well by using the DoF properties of the paired images, which indicates that the training dataset with a limited type of display is sufficient for detection.

5.3 Experiment with Training of Multiple Displays

In this experiment, we utilize two types of displays to train the neural network, while the remaining type of display is considered as the test dataset. Thus, we can validate the robustness of the algorithm in the complex environments due to various models of displays. As shown in Table 4, our algorithm achieves the state-of-the-art performance for every combination of the dataset. Interestingly, despite that the display artifacts of the projector are vastly different from the

other two types of displays, the proposed framework shows the highest accuracy over 93%, which is superior to other algorithms with accuracy under 68%.

5.4 Ablation Study

To validate the roles of the proposed loss terms in Eq. 6, we conduct several ablation studies. First, we reformulate the entire loss of Eq. 6 to consider the DoF center loss and the DoF background loss separately as follows:

$$\mathcal{L}(\mathbf{I}_c, \mathbf{I}_b) = \lambda_{dual}\mathcal{L}_{dual}(\mathbf{I}_c, \mathbf{I}_b) + \lambda_{dof-c}\mathcal{L}_{dof-c}(\mathbf{I}_c) + \lambda_{dof-b}\mathcal{L}_{dof-b}(\mathbf{I}_b), \quad (6)$$

where λ_{dof-b} and λ_{dof-c} control the scales of $\mathcal{L}_{dof-c}$ and $\mathcal{L}_{dof-b}$, respectively. With the reformulated equation, we can investigate the individual effect for $\mathcal{L}_{dof-c}$ and $\mathcal{L}_{dof-b}$. When one of the three loss terms is missing, the scale factor corresponding to the missing loss term is set to 0, while the other factors are set to 0.5, respectively. When two loss terms are missing, only the scale factor for the remaining loss term is set to 1, and the other factors become 0.

The first ablation study evaluates the variation of performance when a part of the three loss terms are missing, through the experiments with the training dataset of a single display as in Section 5.2.

Table 5: Ablation study with training of single display

Train Display	Objectives			5K LCD		WQHD LED		Projector	
	$\mathcal{L}_{dual}$	$\mathcal{L}_{dof-c}$	$\mathcal{L}_{dof-b}$	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
5K LCD	✓			95.54	92.96	89.11	86.78	65.28	62.05
		✓		58.93	58.93	55.34	55.34	50.00	50.00
			✓	69.64	66.00	61.39	58.73	22.83	43.96
	✓	✓		82.14	76.74	71.29	66.98	29.35	49.02
	✓		✓	90.18	85.71	83.17	78.70	52.17	51.18
		✓	✓	58.93	58.93	54.46	54.46	45.65	47.73
	✓	✓	✓	**96.99**	**94.37**	**92.48**	**90.55**	**84.85**	**82.84**
WQHD LED	✓			64.27	73.52	98.02	96.49	94.57	92.58
		✓		58.93	58.93	54.46	54.46	50.00	50.00
			✓	58.93	71.37	**100.00**	**100.00**	82.61	82.61
	✓	✓		52.68	66.23	99.01	98.21	95.65	94.61
	✓		✓	63.39	74.49	**100.00**	**100.00**	92.39	92.39
		✓	✓	68.75	65.35	64.36	60.44	58.70	54.78
	✓	✓	✓	**90.98**	**90.38**	98.48	98.53	**99.25**	**98.53**
Projector	✓			37.50	58.93	64.36	67.31	96.74	93.88
		✓		58.93	58.93	54.46	54.46	50.00	50.00
			✓	41.96	59.55	45.54	54.46	**100.00**	**100.00**
	✓	✓		43.75	60.80	81.19	**84.27**	**100.00**	**100.00**
	✓		✓	41.07	58.93	45.54	54.46	**100.00**	**100.00**
		✓	✓	49.11	60.62	73.26	70.80	88.04	80.70
	✓	✓	✓	**62.41**	**62.29**	**81.95**	80.85	98.48	97.06

Table 5 represents the results of the first ablation study. From the results, we can confirm that the proposed framework considering all the loss terms construct the most general model across various types of displays. Interestingly, the model only with $\mathcal{L}_{dual}$ shows superior performance than the models without $\mathcal{L}_{dof-c}$ or $\mathcal{L}_{dof-b}$, which validates the complementary relationship between the two loss terms in $\mathcal{L}_{dof}$. In addition, the performance declines dramatically when $\mathcal{L}_{dual}$ is missing, which verifies the importance of the paired images to improve the generality of our model.

In the second ablation study, we perform the experiment where all displays are considered for both training and test phases. In this ablation study, we can validate the effect of the three loss terms when the display artifacts in the

training dataset also appear in the test phase. As listed in Table 6, interestingly, even when the framework ignores $\mathcal{L}_{dual}$ and $\mathcal{L}_{dof-c}$, the performance does not decrease much from the full framework because the artifacts can be trained only by $\mathcal{L}_{dof-b}$. Although the performance with L_{dof-b} seems stable when trained with all displays, it declines dramatically when tested with unseen displays. The results validate that the three losses are necessary to ignore the display artifacts.

As indicated in the bottom three rows of Table 6, we conduct additional experiments to demonstrate the effectiveness of our framework. *Stacking Feature* concatenates the two features and masks the unused feature by 0, which validates that our scheme of feature summation is superior than the feature concatenation method. *Color Augmentation* considers the additional augmentation method for the color variation, and the consistent results validate that color inconsistency does not affect the classification accuracy. Finally, we validate the effectiveness of our adversarial training scheme with the shared backbone based on the experiments of separating the backbone, which is listed as *Separated Backbone*. Through various ablation studies, we can validate the effectiveness of the proposed framework for DoF-based image forgery detection.

Table 6: Ablation study with every display

$\mathcal{L}_{dual}$	$\mathcal{L}_{dof-c}$	$\mathcal{L}_{dof-b}$	Acc.	A.P.
✓			89.80	84.16
	✓		54.46	54.46
		✓	90.10	84.53
✓	✓		95.05	91.61
✓		✓	89.77	84.13
	✓	✓	54.46	54.46
✓	✓	✓	**96.04**	**93.17**
Stacking Feature			90.70	84.39
Color Augmentation			94.23	89.66
Separated Backbone			89.95	83.33

6 Conclusion

Recently, online transactions have had an exponential growth and expanded to various applications from e-commerce payments to managing financial accounts from mobile phones. Despite the effort of many companies requesting the usage of their own camera applications for submission of images for online transactions, it is difficult to restrict users from taking a picture of a screen displaying objects, instead of real objects. To detect such cases, we introduce a novel approach utilizing paired images with different depth of field (DoF). In contrast to the flat display panel, the target objects in the real environment are located at various focal lengths from the camera lens, creating difference in DoF in captured images. By utilizing this difference, we can distinguish between the real images and the display images. We introduce a new dataset with 2,752 pairs of images capturing real objects or displayed objects on various types of displays. It is the largest real dataset employing DoF with multi-focus. We also propose a new framework for detecting forged images to focus on the difference of DoF in paired images while avoiding learning individual display artifacts. With numerous ablation studies, we validate that our newly proposed framework achieves the state-of-the-art performance using various displays, including those unseen during training.

References

1. Foran, P.: This rental listing scam is on the rise and catching people off guard (2020, *https://toronto.ctvnews.ca/this-rental-listing-scam-is-on-the-rise-and-catching-people-off-guard-1.4995168* (2020-6-22))
2. Marcellin, F.: Tackling rail fraud in the uk (2020, *https://www.railway-technology.com/features/rail-fraud-in-the-uk/* (2020-1-28))
3. Nguyen, T.T., Nguyen, C.M., Nguyen, D.T., Nguyen, D.T., Nahavandi, S.: Deep learning for deepfakes creation and detection. arXiv preprint arXiv:1909.11573 (2019)
4. Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., Verdoliva, L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv (2018)
5. Zhang, X., Karaman, S., Chang, S.: Detecting and simulating artifacts in gan fake images. In: IEEE International Workshop on Information Forensics and Security. (2019) 1–6
6. Durall, R., Keuper, M., Keuper, J.: Watch your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions. In: IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, United States (2020)
7. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. arXiv preprint arXiv:2003.08685 (2020)
8. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot...for now. In: IEEE Conference on Computer Vision and Pattern Recognition. (2020)
9. Company, S.F.M.A.I.: State farm ® mobile app (2020) `https://www.statefarm.com/customer-care/download-mobile-apps/state-farm-mobile-app`, Last accessed on 2020-7-7.
10. Metz, J.: How to file a car insurance claim from your couch (2020, *https://www.forbes.com/advisor/car-insurance/virtual-claims/* (2020-5-8))
11. Smith, R.: Allstate to move away from physical inspections (2017, *https://www.insurancebusinessmag.com/us/news/breaking-news/allstate-to-move-away-from-physical-inspections-66880.aspx/* (2017-5-5))
12. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations. (2018)
13. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations. (2019)
14. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision. (2017)
15. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2018)
16. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: IEEE Conference on Computer Vision and Pattern Recognition. (2020)

17. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition. (2019) 4401–4410
18. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. CoRR **abs/1912.04958** (2019)
19. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: IEEE International Conference on Acoustics, Speech and Signal Processing. (2019) 8261–8265
20. Li, Y., Chang, M., Lyu, S.: In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). (2018) 1–7
21. Ciftci, U.A., Demir, I.: Fakecatcher: Detection of synthetic portrait videos using biological signals. arXiv preprint arXiv:1901.02212 (2019)
22. Kee, E., Farid, H.: Exposing digital forgeries from 3-d lighting environments. In: IEEE International Workshop on Information Forensics and Security, IEEE (2010) 1–6
23. Carvalho, T., Farid, H., Kee, E.R.: Exposing photo manipulation from user-guided 3d lighting analysis. In: Media Watermarking, Security, and Forensics 2015. Volume 9409., International Society for Optics and Photonics (2015) 940902
24. Peng, B., Wang, W., Dong, J., Tan, T.: Improved 3d lighting environment estimation for image forgery detection. In: IEEE International Workshop on Information Forensics and Security (WIFS), IEEE (2015) 1–6
25. Peng, B., Wang, W., Dong, J., Tan, T.: Optimized 3d lighting environment estimation for image forgery detection. IEEE Transactions on Information Forensics and Security **12** (2016) 479–494
26. Ye, S., Sun, Q., Chang, E.C.: Detecting digital image forgeries by measuring inconsistencies of blocking artifact. In: IEEE International Conference on Multimedia and Expo, Ieee (2007) 12–15
27. Tralic, D., Petrovic, J., Grgic, S.: Jpeg image tampering detection using blocking artifacts. In: International Conference on Systems, Signals and Image Processing, IEEE (2012) 5–8
28. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting World Leaders Against Deep Fakes. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, IEEE (2019) 8
29. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: IEEE Winter Applications of Computer Vision Workshops, IEEE (2019) 83–92
30. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2019)
31. Montserrat, D.M., Hao, H., Yarlagadda, S., Baireddy, S., Shao, R., Horváth, J., Bartusiak, E., Yang, J., Güera, D., Zhu, F., et al.: Deepfakes detection with automatic face weighting. arXiv preprint arXiv:2004.12027 (2020)
32. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: ACM Workshop on Information Hiding and Multimedia Security. (2016) 5–10
33. Kirchner, M.: Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue. In: ACM workshop on Multimedia and security. (2008) 11–20
34. Huang, D.Y., Huang, C.N., Hu, W.C., Chou, C.H.: Robustness of copy-move forgery detection under high jpeg compression artifacts. Multimedia Tools and Applications **76** (2017) 1509–1530

35. Marra, F., Gragnaniello, D., Verdoliva, L., Poggi, G.: Do gans leave artificial fingerprints? In: IEEE Conference on Multimedia Information Processing and Retrieval, IEEE (2019) 506–511
36. Bappy, J.H., Simons, C., Nataraj, L., Manjunath, B., Roy-Chowdhury, A.K.: Hybrid lstm and encoder–decoder architecture for detection of image forgeries. IEEE Transactions on Image Processing **28** (2019) 3286–3300
37. Durall, R., Keuper, M., Pfreundt, F.J., Keuper, J.: Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686 (2019)
38. Demers, J.: Depth of field: A survey of techniques. Gpu Gems **1** (2004) U390
39. Wu, Y., Rivenson, Y., Zhang, Y., Wei, Z., Günaydin, H., Lin, X., Ozcan, A.: Extended depth-of-field in holographic imaging using deep-learning-based autofocusing and phase recovery. Optica **5** (2018) 704–710
40. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
41. Guo, X., Nie, R., Cao, J., Zhou, D., Mei, L., He, K.: Fusegan: Learning to fuse multi-focus image via conditional generative adversarial network. IEEE Transactions on Multimedia **21** (2019) 1982–1996
42. Zhang, J., Liao, Q., Liu, S., Ma, H., Yang, W., Xue, J.h.: Real-mff dataset: A large realistic multi-focus image dataset with ground truth. arXiv preprint arXiv:2003.12779 (2020)
43. Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.: Focusing attention: Towards accurate text recognition in natural images. In: Proceedings of the IEEE international conference on computer vision. (2017) 5076–5084
44. Nejati, M., Samavi, S., Shirani, S.: Multi-focus image fusion using dictionary-based sparse representation. Information Fusion **25** (2015) 72–84
45. Liu, Y., Chen, X., Peng, H., Wang, Z.: Multi-focus image fusion with a deep convolutional neural network. Information Fusion **36** (2017) 191–207
46. Ma, H., Zhang, J., Liu, S., Liao, Q.: Boundary aware multi-focus image fusion using deep neural network. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE (2019) 1150–1155
47. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
48. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323** (1986) 533–536
49. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2014)