Pattern Recognition Letters 181 (2024) 99–105

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A guided-based approach for deepfake detection: RGB-depth integration via features fusion

Giorgio Leporoni, Luca Maiano *, Lorenzo Papa, Irene Amerini

Sapienza University of Rome, Rome, Italy

ARTICLE INFO

Editor: Maria De Marsico

Keywords:
Computer vision
Deepfake
RGBD
Feature fusion

ABSTRACT

Deep fake technology paves the way for a new generation of super realistic artificial content. While this opens the door to extraordinary new applications, the malicious use of deepfakes allows for far more realistic disinformation attacks than ever before. In this paper, we start from the intuition that generating fake content introduces possible inconsistencies in the depth of the generated images. This extra information provides valuable spatial and semantic cues that can reveal inconsistencies facial generative methods introduce. To test this idea, we evaluate different strategies for integrating depth information into an RGB detector and we propose an attention mechanism that makes it possible to integrate information from depth effectively. In addition to being more accurate than an RGB model, our *Masked Depthfake Network* method is +3.2% more robust against common adversarial attacks on average than a typical RGB detector. Furthermore, we show how this technique allows the model to learn more discriminative features than RGB alone.

1. Introduction

Advances in generative techniques allow us to generate artificial images and videos of incredible realism. Many of these contents, such as photos of Pope Francis in sports gear or Donald Trump in handcuffs [1], have gone worldwide in recent months alone. If, until recently, most advanced techniques made us imagine the impossible, today, the application of these technologies in everyday life content is becoming real.

As with any new technology, these advances introduce exciting new applications and dangers. In this work, we focus on this second aspect, proposing a deepfake detection technique based on estimating the depth of the subjects' faces in a video. In particular, our intuition starts from the idea that the generation process introduces 3D inconsistencies in the video, allowing us to identify information that enables the recognition of fake content. This work builds on our previous study [2], proposing substantial improvements in terms of the robustness of the model with respect to adversarial attacks. Unlike the previous work, which limited itself to fusing depth with an early fusion strategy, in this paper, we analyze different depth and RGB fusion methodologies, proposing a solution that integrates an attention mechanism with a late fusion of the two modalities which is able to achieve more accurate estimation performances.

Current deepfake detectors still have several limitations to overcome [3]. First, they tend to overfit training data, resulting in a performance drop that can be very relevant to new attacks. In addition, the more robust detectors are often difficult to interpret, which poses a reliability problem. Moreover, existing state-of-the-art deepfake detection systems rely on neural network-based classification models, which are known to be vulnerable to adversarial examples [4–6]. Depth information provides valuable spatial and semantic cues that can reveal inconsistencies introduced by facial manipulation methods. However, it is unclear to what extent this additional information could contribute to the development of a more robust detector than the corresponding methods based on RGB features alone. This paper shows how depth integration can help mitigate some of these issues. In particular, we show how the attention mechanism proposed in this work produces more interpretable activations than other approaches. Furthermore, we subject our model to a series of adversarial attacks and show that depth integration makes the model more robust to these attacks.

The main contributions of this work are three. (1) We analyze different fusion methods of the RGB and depth channels and propose various experiments to understand the best way to integrate this extra information into a detector. (2) We compare the heatmaps of the proposed model with a model trained only on RGB features and show that integrating the depth with RGB helps the model learn more interpretable and discriminative features with respect to the RGB-only counterpart. (3) We test the robustness of the model against the most commonly used attacks in deepfakes.

* Corresponding author.
E-mail addresses: leporoni.1944533@studenti.uniroma1.it (G. Leporoni), maiano@diag.uniroma1.it (L. Maiano), papa@diag.uniroma1.it (L. Papa), amerini@diag.uniroma1.it (I. Amerini).

https://doi.org/10.1016/j.patrec.2024.03.025
Received 29 June 2023; Received in revised form 7 March 2024; Accepted 27 March 2024
Available online 1 April 2024

The remainder of this paper is organized as follows. Section 2 gives an overview of the state of the art and compares our methodology to existing ones. In Section 3, we introduce the proposed method, which we call *Masked Depthfake Network* (MDN). Section 4 contains the implementation details. Finally, in Section 5, we report our experiments, and in Section 6, we conclude this work.

2. Related work

This section provides an overview of the state of the art for deepfake detection [7,8]. Across several attempts to solve the problem, we can group the main techniques into three large families. (1) *Physiological signals-based* methods look for inconsistencies concerning biological signals. (2) The *identity-based* methods identify irregularities with respect to the target subjects, remodeling the problem as a problem of identification of a subject. Finally, (3) *Learned features* methods detect manipulations as anomalies with respect to the characteristics learned at training time. Our contribution falls into the latter category.

Physiological signals. In the generation phase, some biological characteristics, such as the blinking of the eyes [9] or the blood flow linked to the heartbeat [10,11], are lost. These works look at specific artifacts of the generated videos related to physiological signals. Ciftci et al. [12] focus on the use of up to six different photoplethysmography signals. Unlike these methods, our work does not exploit biological signals but the semantic information of the scene. Even if we apply our method to the specific case of fake face recognition, we believe that an advantage of our approach compared to this category is that it can also be applied to other types of scenes in which the subject is an object or an animal.

Identity-based features. This family of methods approaches the detection problem as a reidentification problem. Agarwal et al. [13] proposed the first approach that falls into this category. It exploits an individual's distinct patterns of facial and head movements to detect fake videos. In later studies, the same research group explored the inconsistencies between the mouth-shape dynamics and a spoken phoneme [14] and proposed an identity-based technique to detect face-swap manipulations [15]. Cozzolino et al. [16] proposed a method that learns temporal facial features, specific to how a person moves while talking, using metric learning coupled with an adversarial training strategy. The advantage is that they do not need any training data for fakes but only train on authentic videos. Moreover, they utilize high-level semantic features, which enable robustness to widespread and disruptive forms of post-processing. Similar to the previous category, a limitation of these approaches is that they are only applicable to contexts in which the video portrays human subjects, while our method, as well as those belonging to the next category, exploit features applicable to generic images.

Learned features. This category includes all methods that use features that can be automatically learned from a model and are not explicitly based on biological characteristics or the identity of a subject. Afchar et al. [17] presented one of the first approaches for deepfake detection based on supervised learning. It focuses on mesoscopic features to analyze the video frames using a network with few layers. Rössler et al. [18] analyzed the performance of several CNN architectures for deepfake video detection and showed that deeper networks are more effective for this task, especially on low-quality video. Zhao et al. [19] explore image deepfake detection by dividing the image into small patches. Each patch gets a consistency value with respect to all the others. Therefore, manipulations can be identified as patches that have lower consistency with respect to others. Dang et al. [20] examine several ways to combine attention mechanisms in CNN networks to highlight tampered image regions and then guide the network in the detection phase. Other studies rely on processing temporal dependencies and patterns between frames so that inconsistencies and anomalies in deepfake videos can be detected. For example, Sabir et al. [21] study how to pair an RNN module with a CNN backbone in an end-to-end model to improve accuracy by keeping frames in a temporal relationship. Caldelli et al. [22] propose a methodology based on optical flow features. In the same direction, Saikia et al. [23] use optical flow and estimate the frame correlation with an LSTM to detect any temporal skew. Similarly, Ismail et al. [24] propose a deepfake detection method based on LSTMs. Recently, a different approach based on face depth maps have been proposed by Maiano et al. [2]. The latter study leverages the use of monocular depth estimation methodologies, i.e., deep learning solutions such as encoder–decoder architectures able to extract a per-pixel distance map from a single input image [25], in order to take advantage of depth map inconsistencies (flatten maps) introduced during the generation phase between real and fake samples. Differently from the previous study, in this work, the depth is precomputed and fused with the RGB information with an early fusion strategy. Different from this approach, in this work, we propose to exploit an attention mechanism and combine it with a late fusion strategy. A similar approach [26] proposes a depth prediction and a triplet feature extraction network. Our method differs from this one in the way in which the depth is fused, thanks to a late fusion strategy and the proposed attention mechanism.

3. Proposed method

Our method is based on the intuition that, as shown in our previous study [2], the deepfake creation process introduces inconsistencies in the depth of the face. Consequently, combining the depth information with the RGB image will improve the learning process and make it more stable and possibly more robust against adversarial attacks. Our goal is, therefore, to understand how RGB and depth information can best be combined to obtain greater accuracy than RGB features on which the networks typically focus. Unlike our previous study, we propose a late fusion mechanism combining the RGB and depth features. Moreover, we propose an attention mechanism that guides the learning of the feature-depth network based on the most important features identified by the RGB one. This allows us to keep the two inputs in two separate streams but, at the same time, direct learning towards a common feature space. The proposed method is composed of the following two steps. (1) First, we extract depth from the whole frame using the pretrained model introduced by Khan et al. [27]. Since we know that image resize tends to destroy fundamental traces for the recognition of fakes, and that manipulations usually focus on the face and the areas around it, we extract the person's face using a $W \times H$ crop. (2) Then, as shown in Fig. 1, we input the RGB and depth patches to the deepfake detection model to classify the frame as *true* or *false*.

More details on both steps are provided in the remainder of this section. Specifically, Section 3.1 explains the preprocessing operations we perform to extract depth and crop the face, and Section 3.2 describes our method of detecting deepfakes.

3.1. Pre-processing

Our method revolves around the depth estimation task. We assume that the deepfake generation process introduces distortions in the depth of the subject's face, which can be crucial for the final classification. For this step, we rely on the method introduced by Maiano et al. [2].

As mentioned above, the proposed pipeline starts with estimating the frame depth. For this purpose, we use FaceDepth [27], a technique for estimating the monocular depth of faces. This network has been specifically trained to calculate the distance of faces from the camera. FaceDepth can detect details of facial features and obtain precise depth information for each facial point. This allows for accurate discrimination of facial features and allows us to estimate the differences between real and fake faces more accurately.

Image resizing can eliminate important information that can help the model in the classification task, so to avoid this kind of problem,

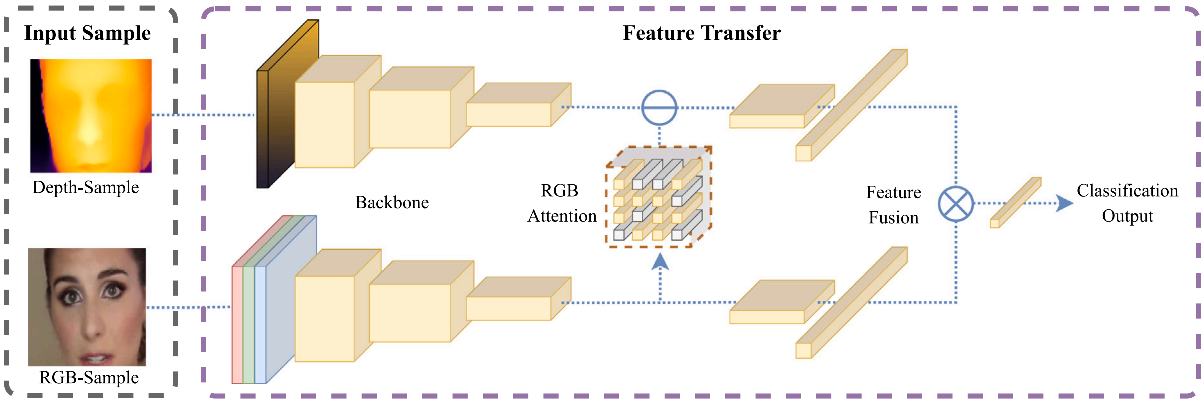

Fig. 1. Our proposed pipeline. RGB and depth characteristics are analyzed separately by two MobileNet v2. We introduce an attention mechanism that masks the less important depth features based on the RGB features. Finally, we merge the features through a concatenation to proceed to the classification.

we extract a crop centered on the subject's face. We use the Dlib[1] library for face detection and extraction.

3.2. Deepfake detection

In this section, we discuss the central part of our contribution. The proposed model takes as input the original RGB patch of size $W \times H \times 3$ and the estimated depth map of size $W \times H \times 1$. Given RGB's larger number of channels, using a single input that concatenates RGB and depth information may not be optimal because RGB information may get more network attention than depth. To avoid this, we have developed an ad hoc architecture, shown in Fig. 1. The whole architecture consists of two different networks, one for each type of input (RGB and depth).

The proposed architecture, which we call *Masked Depth Network (MDN)*, comprises two parallel networks. The RGB network is designed to process individual RGB frames and extract relevant features for deepfake detection. In contrast, the depth network captures depth-related inconsistencies that are often difficult to eliminate in deepfake videos. The information extracted from the two networks is merged before classification by concatenating the output of the last convolutional layer of both streams. This allows us to integrate the information extracted from the RGB and depth networks while preserving the most discriminating aspects of both channels.

In addition to the fusion phase, we introduce an *attention mechanism* to guide the depth network in selecting the most essential features. This step enforces the fusion process by highlighting regions of interest for deepfake detection. The attention is introduced by masking the weights of the RGB network and integrating this mask into the depth network. Formally, given a weight matrix W, we compute the attention mask $a(w_i)$ for all $w_i \in W$ as follows.

$$a(w_i) = \begin{cases} 0, & \text{if } w_i < 0. \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

In our experiments, we apply this masking operation on the fourth convolutional block of the MobileNet v2 [28] architecture. This attention mechanism allows the depth network to dynamically adjust its attention to focus on the most informative regions indicated by the RGB stream and which are expected to contain critical depth-based cues for distinguishing real video from deepfakes. As shown in Section 5.1, this architectural choice helps to achieve better performance than simple feature concatenation.

In summary, our proposed architecture's fusion and mask generation steps enable the integration of RGB and depth channels while selectively highlighting informative regions. As we will discuss in Section 5, this approach improves the architecture's overall accuracy and robustness, leverages both channels' strengths, and facilitates the extraction of relevant features for effective deepfake detection. We incorporate augmentation techniques on the pre-trained model during the training process to address the overfitting problem and improve our architecture's generalization capability. The implementation details are discussed in the next section.

4. Implementation details

The proposed method has been implemented using PyTorch[2] deep learning API. The trained architectures are initialized on ImageNet pretrained weights and trained with a CrossEntropy loss function for 30 epochs with a batch size of 192 using Adam optimizer [29] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate of 0.0001. Moreover, to improve the generalization performances of trained models, data augmentation has been incrementally performed during the training epochs, i.e, by increasing the augmentation effect with the increase of the number of epochs. We examine multiple augmentation strategies and transformations as proposed in [30–33]. We evaluate our method on commonly used evaluation metrics such as estimation accuracy, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC) on the FaceForensics++ dataset [18]. This dataset is widely used in the state of the art due to its heterogeneity and complexity. The dataset consists of more than 1000 YouTube videos containing real and manipulated faces in various settings and conditions. The manipulated videos were generated using various deepfake techniques, such as facial reenactment, face swapping, and expression manipulation. The dataset consists of four classes of forgery attacks: (1) *Deepfakes* (DF), (2) *Face2Face* (F2F), (3) *FaceSwap* (FS), and (4) *NeuralTextures* (NT). In our experiments, we report results on individual classes and the entire test set of all four classes. We refer to this last case indicating it as *ALL*. Moreover, the dataset has different compression levels for each video, namely RAW (uncompressed), C23, and C40. Due to space constraints, we report only the experiments on the two limiting cases, namely RAW and C40.

5. Results

This section shows the effectiveness of including depth information for deepfake detection versus a standard RGB approach. Precisely, we chose the MobileNet v2 [28] as the backbone in all experiments, which

[1] http://dlib.net/.

[2] Code and corresponding pre-trained weights are made publicly available at the following GitHub repository: https://github.com/gleporoni/rgbd-depthfake.

Table 1
Quantitative results obtained on deepfake detection task for RAW and C40 dataset settings when trained on Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT), and all (ALL) forgeries. The best results for each configuration are reported in bold.

Class	Testing Set (RAW)				
	RGB	D	EF [2]	LF	MDN
DF	96,00%	89,23%	95,35%	96,59%	**96,86%**
F2F	95,35%	84,75%	95,38%	95,57%	**95,85%**
FS	95,32%	81,23%	95,62%	**96,33%**	96,29%
NT	92,01%	78,77%	92,30%	**92,76%**	92,65%
ALL	95,02%	83,23%	**95,09%**	94,81%	94,87%
Class	Testing Set (C40)				
	RGB	D	EF [2]	LF	MDN
DF	88,15%	73,46%	88,65%	90,75%	**91,26%**
F2F	**82,57%**	66,39%	82,13%	82,25%	81,82%
FS	86,11%	67,00%	85,45%	86,73%	**87,17%**
NT	70,77%	59,26%	70,77%	**71,00%**	70,50%
ALL	82,37%	79,59%	82,09%	82,25%	**82,43%**

Table 2
AUC values obtained on deepfake detection task for RAW and C40 dataset settings when trained on Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT), and all (ALL) forgeries. The best results for each configuration are reported in bold.

Class	Testing Set (RAW)				
	RGB	D	EF [2]	LF	MDN
DF	99,08%	95,73%	99,22%	99,02%	**99,33%**
F2F	99,07%	90,65%	**99,19%**	99,09%	98,88%
FS	99,09%	85,20%	99,19%	99,50%	**99,51%**
NT	97,52%	85,38%	97,72%	**97,91%**	97,90%
ALL	98,29%	78,33%	98,25%	98,37%	**98,50%**
Class	Testing Set (C40)				
	RGB	D	EF [2]	LF	MDN
DF	93,79%	77,62%	93,79%	96,71%	**96,83%**
F2F	89,51%	68,01%	89,01%	**90,21%**	90,15%
FS	90,48%	69,71%	90,72%	94,17%	**94,34%**
NT	74,32%	58,75%	73,06%	**77,70%**	77,55%
ALL	78,19%	54,50%	78,76%	80,16%	**81,20%**

Table 3
AUC values by our proposed MDN compared to state-of-the-art methods. The best two are shown in bold and underlined respectively.

Class	Type	ALL (RAW)
DR [11]	Physiological signals	98,00%
PPG [10]		93,50%
AB [14]	Identity-based	97,00%
IID [36]		**99,00%**
EF [2]	Learned features	95,09%
DGN [26]		98,30%
MIL [37]		97,73%
FKSPT [38]		<u>98,50%</u>
MDN (our)		<u>98,50%</u>

is demonstrated to perform well despite being a lightweight architecture [2]. In addition, to leverage the effectiveness of the proposed attention mechanism included in our Masked Depthfake Network and inspired by fusion strategies discussed by Ophoff et al. [34] and Zhou et al. [35], we also compare the proposed method with different architectural and input configurations. Precisely, to validate the proposed architecture, we introduce four baseline structures where we modify how the RGBD input is provided to the model. Below is a detailed description of each model.

- *RGB*: it consists of a single MobileNet v2 network that is trained on the RGB frames.
- *Depth (D)*: it consists of a single MobileNet v2 network trained only on depth maps.
- *Early fusion (EF)*: in this scenario, the RGB and depth inputs are stacked and passed to the network as a single (4-channel) input as done in Maiano et al. [2]. The addition of the depth channel beside RGB creates the need to use correct weights initialization for pre-trained models. To do this, we average the weights of the first layer of the RGB network model trained on ImageNet.
- *Late fusion (LF)*: it comprises two separate MobileNet v2 networks whose output features are concatenated into a single vector before the classification layers. The combined vector is then passed to the fully connected layers for the final classification phase.

The remainder of this section is organized as follows. We first compare the deepfake recognition performance of the proposed model with the baselines described above. Then, we examine the activation maps of the proposed model against the RGB baseline to see if the addition of depth and the proposed attention system lead the model to pay attention to more discriminating features. Finally, we conclude the section by studying the robustness of the proposed method against common *black-box* adversarial attacks for deepfakes.

5.1. Detection performance

Our first analysis aims to quantitatively demonstrate the effectiveness of the addition of the depth channel to standard RGB approaches and validate the proposed Masked Depthfake Network architecture with respect to the baselines. Through these experiments, the intent is to understand what is the most effective way to use depth for this task. Table 1 shows the overall accuracy obtained at testing time over the different forgeries and compressions levels of the FaceForensic++ dataset. The ROC curves and their respective AUC (area under the curve) values are reported in Figs. 2, 3 and Table 2 respectively.

The results show that integrating the depth channel into the standard RGB approach guarantees increased detection performance. The MDN and the LF usually perform better than the other baselines for all types of forgeries and compressions. These architectures achieve an average accuracy and AUC boost of up to +1.01% and +0.42% on the RAW dataset and up to +3.11% and +3.86% on the C40 dataset respectively, demonstrating that the depth information combined with RGB one is able to improve the overall detection process. Moreover, we can notice that the RGB network outperforms the D network alone for both RAW and C40 datasets by an average percentage of 12.07% and of 15.55% on the AUC. This is perfectly explained by the fact that, besides having fewer channels, the depth information is estimated starting from the RGB. However, when we combine the two pieces of information, the results confirm the hypothesis that depth information helps the model better discriminate between fake and real examples based on inconsistencies in image depth.

Finally, in Table 3 we compare our results against different state-of-the-art methods. DeepRhythm [11], and the PPG-based method from Ciftci et al. [10] use physiological signals. The appearance and behavior (AB) method from Agrawal et al. [14] and the Implicit Identity Driven Deepfake Face Swapping Detection (IID) [36] method are identity-based methods. Finally, the Depth Map-guided Triplet Network [26] (DGN), the Multiple Instance Networks (MIL [37]), and Fakespotter (FKSPT [38]) use learned features similar to our proposed method. Our proposed method is the best runner-up after IID, which performs slightly better than our method (+0.5%). These results confirm the contribution introduced by depth compared to other methodologies. In the next section, we delve further into the contribution of depth to identify any limitations.

5.2. Feature analysis

We now analyze the activation maps of the proposed Masked Depthfake Network from a qualitative perspective. In particular, with this analysis, we want to understand if the attention mechanism leads

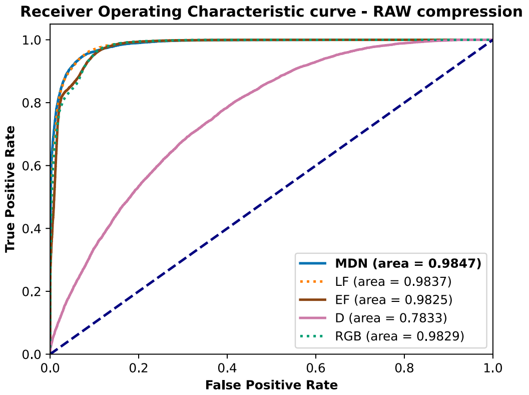

Fig. 2. ROC Curve results obtained on deepfake detection task for RAW dataset when trained on all (ALL) forgeries.

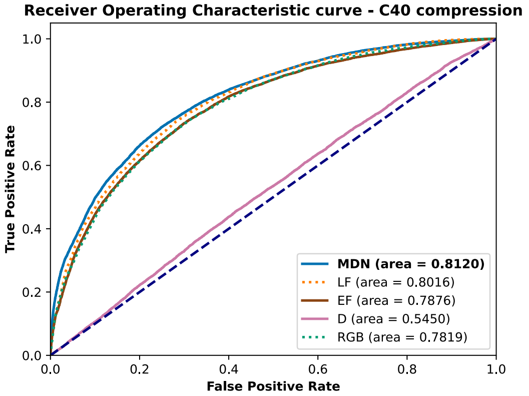

Fig. 3. ROC Curve results obtained on deepfake detection task for C40 dataset when trained on all (ALL) forgeries.

the network to focus on more discriminative and, therefore, more interpretable features than the RGB counterpart. Consequently, we calculate and display the activation maps of the last convolutional layer of the Masked Depthfake Network and the RGB baseline using the GradCam [39] method. We report an example of the obtained *Real* and *Fake* output heatmaps in Fig. 4.

The first row of the figure represents the RGB and corresponding depth inputs extracted from the FaceForensic++ dataset (RAW). The second and third rows report the heatmaps of the RGB and depth models, respectively. The heatmaps show that the RGB model produces more or less uniform activations, which does not give us particular indications on any area of the face. This could suggest that the model may have overfitted the training samples, making it less robust and interpretable. In the case of depth, the activation is most robust in the area around the nose. In the EF model, we notice that this difference between RGB and depth disappears, highlighting the problem of immediately merging the necessary features. The effect of early fusion is to reduce the depth contribution compared to other fusion methods. The advantage of late fusion becomes evident for the LF and MDN models, where the depth and RGB components have different turn-ons. In particular, we can observe a stronger activation of the MDN in correspondence with the nose and eyes area, which shows how the attention mechanism manages to concentrate the model's attention towards the places most subject to manipulation. This also highlights a possible limitation of this approach: the model's performance is strictly linked to the accuracy of the depth estimation model. However, this problem can be easily mitigated with a more accurate depth estimation method.

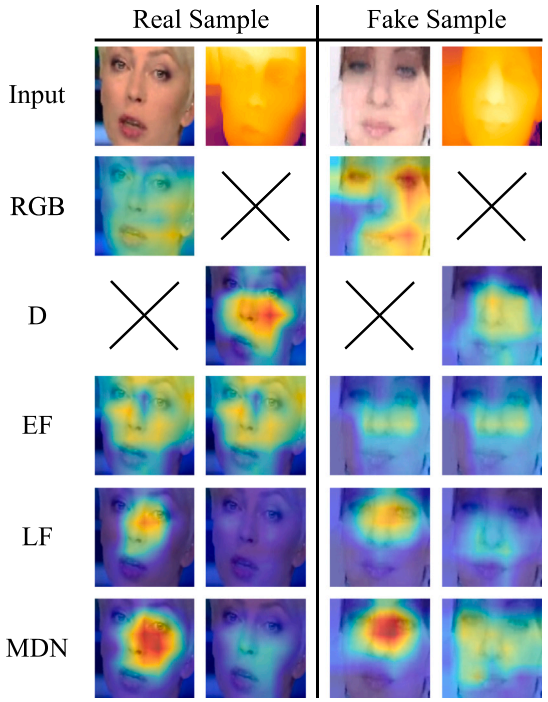

Fig. 4. Heatmaps generated by the GradCam algorithm. The CAM RGB shows the obtained heatmaps for the RGB baseline model, while CAM MDN for the proposed method.

Differently, we can notice that the MDN network activates on specific regions of the face. Precisely, the model focuses on the nose region for the authentic RGB image, while for the fake one, the model focuses on the nose and the eyes. Similarly, but with an opposite behavior, we notice that the depth network focuses on particular interest areas of the mouth, nose, and eyes.

Summarizing, we can conclude that although the RGB approach achieves good results in terms of accuracy, it does not pay particular attention to specific regions of the face. Indeed, it also takes into consideration the background areas, and this is why this approach is, in general, less reliable. This could suggest that instead of learning facial features, the network is overfitting the dataset, storing general information only partially related to the characteristics introduced by the generation process. Contrarily, our proposed method, which focuses on the tampered region of the image, leads to improved results in terms of accuracy and makes the whole pipeline more robust, as discussed in the next section.

5.3. Robustness to adversarial attacks

To have a complete overview of the proposed method against the RGB baseline, we study its robustness against adversarial attacks. An attacker may decide to introduce imperceptible disturbances in the fake video to bypass deepfake detectors. Specifically, we test the robustness of the models against *black-box* attacks discussed in Gandhi and Jain, and Hussain et al. [40,41]. Since these attacks include *Guassian blur* (BLR), *Guassian noise* (NSE), *rescaling* (RSC), and *translation* (TRN), which are also used in the data augmentation strategy, for a correct comparison, we use the RGB and MDN methods trained without any data augmentation strategy.

Tables 4 and 5 report the performance of all the baseline models against every single attack as well as their combination (CMB). Based on the obtained results, we can notice that all the RGBD approaches are able to outperform the standard RGB one in almost all of the experiments. More in detail, in the case of the RAW dataset (see Table 4), the MDN achieves an averaged percentage boost of +3.89%, +2.54%, and +0.63% with respect to the RGB, EF, and LF methods respectively.

Table 4
Accuracy results obtained on deepfake detection task for RAW dataset settings when Blur (BLR), Noise (NSE), Rescale (RSC), Translation (TRN), and all Combined (CMB) black box attacks are applied. The best results are in bold and the second best are underlined.

Attack	Model	Testing Set (RAW)				
		DF	F2F	FS	NT	ALL
BLR	RGB	50,32%	54,30%	50,45%	49,00%	79,64%
	D	51,47%	64,84%	**51,69%**	49,39%	**79,75%**
	EF [2]	**53,05%**	51,63%	50,60%	49,58%	77,50%
	LF	50,34%	69,00%	50,47%	**51,92%**	78,07%
	MDN	50,98%	**70,38%**	50,62%	50,73%	79,71%
NSE	RGB	85,80%	88,73%	93,83%	76,21%	92,06%
	D	50,46%	51,88%	60,41%	50,64%	29,03%
	EF [2]	95,32%	95,36%	95,61%	**92,27%**	**95,08%**
	LF	95,58%	94,96%	**95,89%**	92,11%	94,98%
	MDN	**95,69%**	**95,43%**	95,67%	91,61%	94,67%
RSC	RGB	60,63%	60,60%	50,58%	53,89%	74,48%
	D	52,23%	68,36%	**52,44%**	49,46%	**79,70%**
	EF [2]	**67,06%**	56,91%	50,58%	54,60%	70,42%
	LF	56,11%	73,91%	50,65%	**62,17%**	74,48%
	MDN	61,62%	**75,64%**	50,62%	60,80%	78,76%
TRN	RGB	95,87%	95,11%	95,19%	91,63%	94,89%
	D	88,16%	81,75%	75,49%	77,57%	82,52%
	EF [2]	95,19%	95,15%	95,42%	91,80%	**94,92%**
	LF	96,42%	95,22%	**96,29%**	**92,52%**	94,76%
	MDN	**96,75%**	**95,49%**	96,23%	92,13%	94,51%
CMB	RGB	50,31%	55,22%	50,33%	49,47%	77,95%
	D	50,80%	53,74%	**51,05%**	49,45%	79,57%
	EF [2]	**52,44%**	51,75%	50,52%	49,64%	77,83%
	LF	50,11%	62,67%	50,51%	**51,33%**	77,63%
	MDN	50,27%	**64,64%**	50,61%	50,46%	**79,80%**

Table 5
Accuracy results obtained on deepfake detection task for C40 dataset settings when Blur (BLR), Noise (NSE), Rescale (RSC), Translation (TRN), and all Combined (CMB) black box attacks are applied. The best results are in bold and the second best are underlined.

Attack	Model	Testing Set (C40)				
		DF	F2F	FS	NT	ALL
BLR	RGB	66,03%	67,82%	58,02%	**56,00%**	79,27%
	D	55,27%	60,35%	57,67%	50,22%	79,69%
	EF [2]	**75,94%**	67,00%	55,94%	53,17%	74,34%
	LF	63,29%	70,57%	69,09%	50,27%	**80,00%**
	MDN	75,17%	**74,60%**	**71,92%**	49,70%	79,90%
NSE	RGB	87,20%	81,05%	85,57%	62,00%	81,41%
	D	61,11%	58,82%	61,74%	59,38%	79,55%
	EF [2]	88,63%	82,15%	85,48%	**70,73%**	**82,09%**
	LF	**90,98%**	**82,38%**	86,65%	70,28%	82,02%
	MDN	89,75%	81,69%	**86,86%**	70,65%	82,05%
RSC	RGB	73,02%	70,19%	64,67%	**57,18%**	80,11%
	D	56,12%	60,25%	58,36%	51,11%	79,74%
	EF [2]	68,64%	69,23%	61,28%	55,40%	75,51%
	LF	**80,76%**	71,58%	73,67%	51,41%	**80,44%**
	MDN	78,37%	**74,96%**	**76,59%**	50,46%	80,38%
TRN	RGB	87,63%	81,23%	84,83%	70,13%	81,07%
	D	71,74%	61,76%	64,71%	57,35%	79,65%
	EF [2]	87,50%	80,80%	84,67%	69,64%	81,37%
	LF	89,72%	**81,84%**	86,14%	**70,73%**	81,68%
	MDN	**89,76%**	81,46%	**86,24%**	69,81%	**81,82%**
CMB	RGB	58,73%	66,32%	55,75%	50,01%	79,67%
	D	50,50%	56,15%	54,80%	50,13%	79,73%
	EF [2]	60,45%	63,68%	54,07%	**51,97%**	73,70%
	LF	67,00%	68,70%	64,32%	49,85%	79,59%
	MDN	**71,96%**	**73,22%**	**65,96%**	50,52%	**79,80%**

Similarly, in the case of the compressed (C40) dataset, reported in Table 5, we can notice that the average improvement achieved by MDN over the RGB, EF, and LF methods is equal to +3.48%, +3.96% and +1.18% respectively.

Based on the reported values, we can conclude that the proposed method could be a viable solution to improve the estimation performances and the robustness against adversarial attacks in the deepfake detection task.

6. Conclusion

This paper has explored different fusion strategies between RGB and depth for deepfake detection. In our experiments, we show how incorporating depth information into the detection, the *Masked Depthfake Network* can improve the accuracy and robustness of deepfake detection systems. This is mainly due to the additional information introduced with the depth channel, which provides valuable spatial cues that are difficult to replicate in synthetic video. By leveraging this additional information, our solution identifies subtle inconsistencies that traditional RGB visual techniques can overlook. Furthermore, depth data has also shown greater resistance to adversary attacks and manipulations, providing a more robust defense against deepfake techniques. A possible limitation of this proposed methodology could be its dependency on the depth estimation model, which, if not sufficiently robust, could introduce semantic errors, reducing the detector's performance. However, using an accurate depth estimation method, this problem can be easily mitigated.

The analyses reported in this article open up new questions for the community to explore. First, verifying the robustness of a depth-based method against other black-box and white-box attacks would be interesting. Furthermore, it would be interesting to analyze depth inconsistencies from a temporal point of view, exploiting both geometric and temporal signals. Finally, this same analysis can be further extended to other datasets and generative techniques.

CRediT authorship contribution statement

Giorgio Leporoni: Conceptualization, Investigation, Methodology, Software, Visualization, Writing – review & editing. **Luca Maiano:** Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Lorenzo Papa:** Conceptualization, Supervision, Writing – original draft. **Irene Amerini:** Supervision, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This study has been partially supported by SERICS, Italy (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU and Sapienza University of Rome, Italy project "EV2" (003_009_22).

References

[1] M. Novak, AI image creator midjourney halts free trials but it has nothing to do with the pope's jacket, 2023, URL: https://www.forbes.com/sites/mattnovak/2023/03/31/ai-image-creator-midjourney-halts-free-trials-but-it-has-nothing-to-do-with-the-popes-jacket (accessed: 2023-06-14).

[2] L. Maiano, L. Papa, K. Vocaj, I. Amerini, DepthFake: A depth-based strategy for detecting deepfake videos, in: J.-J. Rousseau, B. Kapralos (Eds.), Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges, Springer Nature Switzerland, Cham, 2023, pp. 17–31.

[3] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, L. Verdoliva, Are GAN generated images easy to detect? A critical analysis of the state-of-the-art, in: 2021 IEEE International Conference on Multimedia and Expo, ICME, 2021, pp. 1–6, http://dx.doi.org/10.1109/ICME51207.2021.9428429.

[4] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, J. McAuley, Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2021, pp. 3348–3357.

[5] Z. Sun, Y. Han, Z. Hua, N. Ruan, W. Jia, Improving the efficiency and robustness of deepfakes detection through precise geometric features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3609–3618.

[6] P. Neekhara, B. Dolhansky, J. Bitton, C.C. Ferrer, Adversarial threats to DeepFake detection: A practical perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 923–932.

[7] I. Amerini, A. Anagnostopoulos, L. Maiano, L.R. Celsi, Deep learning for multimedia forensics, Found. Trends® Comput. Graph. Vis. 12 (4) (2021) 309–457, http://dx.doi.org/10.1561/0600000096.

[8] L. Verdoliva, Media forensics and DeepFakes: An overview, IEEE J. Sel. Top. Sign. Proces. 14 (5) (2020) 910–932, http://dx.doi.org/10.1109/JSTSP.2020.3002101.

[9] T. Jung, S. Kim, K. Kim, DeepVision: Deepfakes detection using human eye blinking pattern, IEEE Access 8 (2020) 83144–83154, http://dx.doi.org/10.1109/ACCESS.2020.2988660.

[10] U.A. Ciftci, İ. Demir, L. Yin, How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals, in: 2020 IEEE International Joint Conference on Biometrics, IJCB, 2020, pp. 1–10, http://dx.doi.org/10.1109/IJCB48548.2020.9304909.

[11] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, J. Zhao, DeepRhythm: Exposing DeepFakes with attentional visual heartbeat rhythms, 2020, arXiv:2006.07634.

[12] U.A. Ciftci, I. Demir, L. Yin, FakeCatcher: Detection of synthetic portrait videos using biological signals, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1, http://dx.doi.org/10.1109/TPAMI.2020.3009287.

[13] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li, Protecting world leaders against deep fakes, in: CVPR Workshops, Vol. 1, 2019, p. 38.

[14] S. Agarwal, H. Farid, T. El-Gaaly, S.-N. Lim, Detecting deep-fake videos from appearance and behavior, in: 2020 IEEE International Workshop on Information Forensics and Security, WIFS, IEEE, 2020, pp. 1–6.

[15] S. Agarwal, H. Farid, O. Fried, M. Agrawala, Detecting deep-fake videos from phoneme-viseme mismatches, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 660–661.

[16] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, L. Verdoliva, Id-reveal: Identity-aware deepfake video detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15108–15117.

[17] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: 2018 IEEE International Workshop on Information Forensics and Security, WIFS, IEEE, 2018, pp. 1–7.

[18] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.

[19] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, W. Xia, Learning self-consistency for deepfake detection, 2021, arXiv:2012.09311.

[20] H. Dang, F. Liu, J. Stehouwer, X. Liu, A. Jain, On the detection of digital face manipulation, 2020, arXiv:1910.01717.

[21] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, 2019, arXiv:1905.00582.

[22] R. Caldelli, L. Galteri, I. Amerini, A. Del Bimbo, Optical Flow based CNN for detection of unlearnt deepfake manipulations, Pattern Recognit. Lett. 146 (2021) 31–37.

[23] P. Saikia, D. Dholaria, P. Yadav, V. Patel, M. Roy, A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features, 2022, arXiv:2208.00788.

[24] A. Ismail, M. Elpeltagy, M.S. Zaki, K. Eldahshan, An integrated spatiotemporal-based methodology for deepfake detection, Neural Comput. Appl. 34 (24) (2022) 21777–21791, http://dx.doi.org/10.1007/s00521-022-07633-3.

[25] Y. Ming, X. Meng, C. Fan, H. Yu, Deep learning for monocular depth estimation: A review, Neurocomputing 438 (2021) 14–33.

[26] B. Liang, Z. Wang, B. Huang, Q. Zou, Q. Wang, J. Liang, Depth map guided triplet network for deepfake face detection, Neural Netw. 159 (2023) 34–42, http://dx.doi.org/10.1016/j.neunet.2022.11.031, URL: https://www.sciencedirect.com/science/article/pii/S0893608022004725.

[27] F. Khan, S. Hussain, S. Basak, J. Lemley, P. Corcoran, An efficient encoder–decoder model for portrait depth estimation from single images trained on pixel-accurate synthetic data, Neural Netw. 142 (2021) 479–491, http://dx.doi.org/10.1016/j.neunet.2021.07.007, URL: https://www.sciencedirect.com/science/article/pii/S0893608021002707.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[29] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[30] L. Chai, D. Bau, S.-N. Lim, P. Isola, What makes fake images detectable? Understanding properties that generalize, 2020, arXiv:2008.10588.

[31] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, L. Verdoliva, Are GAN generated images easy to detect? A critical analysis of the state-of-the-art, 2021, arXiv:2104.02617.

[32] S. Das, S. Seferbekov, A. Datta, M.S. Islam, M.R. Amin, Towards solving the DeepFake problem : An analysis on improving DeepFake detection using dynamic face augmentation, 2021, arXiv:2102.09603.

[33] H. Huang, A. Geiger, D. Zhang, GOOD: Exploring geometric cues for detecting objects in an open world, 2023, arXiv:2212.11720.

[34] T. Ophoff, K.V. Beeck, T. Goedemé, Exploring RGBDepth fusion for real-time object detection, Sensors 19 (4) (2019) 866, http://dx.doi.org/10.3390/s19040866.

[35] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, L. Shao, RGB-D salient object detection: A survey, Comput. Vis. Media 7 (1) (2021) 37–69, http://dx.doi.org/10.1007/s41095-020-0199-z.

[36] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, D. Ye, Implicit identity driven deepfake face swapping detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4490–4499.

[37] X. Wang, Y. Yan, P. Tang, X. Bai, W. Liu, Revisiting multiple instance neural networks, Pattern Recognit. 74 (2018) 15–24, http://dx.doi.org/10.1016/j.patcog.2017.08.026, URL: https://www.sciencedirect.com/science/article/pii/S0031320317303382.

[38] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, Y. Liu, Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces, 2019, arXiv preprint arXiv:1909.06122.

[39] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2019) 336–359, http://dx.doi.org/10.1007/s11263-019-01228-7.

[40] A. Gandhi, S. Jain, Adversarial perturbations fool deepfake detectors, 2020, arXiv:2003.10596.

[41] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, J. McAuley, Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples, 2020, arXiv:2002.12749.