Inconsistency-Aware Wavelet Dual-Branch Network for Face Forgery Detection

Gengyun Jia, Meisong Zheng, Chuanrui Hu, Xin Ma, Yuting Xu, Luoqi Liu, Yafeng Deng, and Ran He, *Senior Member, IEEE*

***Abstract*—Current face forgery techniques can generate high-fidelity fake faces with extremely low labor and time costs. As a result, face forgery detection becomes an important research topic to prevent technology abuse. In this paper, we present an inconsistency-aware wavelet dual-branch network for face forgery detection. This model is mainly based on two kinds of forgery clues called inter-image and intra-image inconsistencies. To fully utilize them, we firstly enhance the forgery features by using additional inputs based on stationary wavelet decomposition (SWD). Then, considering the different properties of the two inconsistencies, we design a dual-branch network that predicts image-level and pixel-level forgery labels respectively. The segmentation branch aims to recognize real and fake local regions, which is crucial for discovering intra-image inconsistency. The classification branch learns to discriminate the real and fake images globally, thus can extract inter-image inconsistency. Finally, bilinear pooling is employed to fuse the features from the two branches. We find that the bilinear pooling is a kind of spatial attentive pooling. It effectively utilizes the rich spatial features learned by the segmentation branch. Experimental results show that the proposed method surpasses the state-of-the-art face forgery detection methods.**

***Index Terms*—Face forgery detection, stationary wavelet decomposition, dual-branch network, bilinear pooling.**

I. Introduction

IMAGE forgery which has been applied for decades for either good or evil purposes is not a new technology. Traditional methods involve a large amount of time and labor costs, and the users need to be skilled in related tools. Many early works propose to detect such image manipulations via blind methods that do not use any data for training. These methods often exploit artifacts generated from specific image processing stages such as photography and compressed transmission. Optical distortion [1], texture patterns [2], [3], noise features [4], [5] and compression artifacts [6], [7] are some of the most popular clues. Recently, deep learning becomes an effective framework to deal with this challenge. Some methods use convolutional neural networks combined with other features such as noise residuals [8].

However, recent development on deep generative learning brought technology revolution into image forgery research. Many powerful methods has been developed [9], [10] and have been successfully applied in many areas [11], [12], [13], [14], [15], [16], [17]. With the help of tools based on these methods, a non-expert is able to forge images and videos better than previous experts in even several seconds. The tools focusing on human faces are especially popular since they have various potential applications. In this situation, technology abuse may cause severely bad influences especially in the online social medias that spread fake information very fast. Therefore, it is very important and urgent to develop effective face forgery detection methods against current face forgery algorithms.

To promote the face forgery detection research, many datasets [18], [19], [20] were proposed recently. These datasets are based on popular algorithms such as DeepFakes [21] and FaceSwap [22]. The imperfect algorithms leave artifacts for us to identify the forged images. We find that the manipulations are always applied in local image regions in these methods. For example, DeepFakes only exchanges and adjust rectangular regions covering the facial features. As a result, the artifacts can be categorized into two aspects. The first is the inter-image inconsistency. It reflects the general differences between real and forged images. The second is the intra-image inconsistency that describes the differences between the forged regions and unforged regions inside one image. We use an example to show the two clues in Figure 1. It can be seen that in the left fake image, the skin colors are different in the two small regions, while the corresponding real image shows a uniform skin appearances. The right two are the second level decomposed images with stationary wavelet decomposition. we can see that there are abnormal high-frequency responses in the fake images, while the corresponding region in the real image is smooth. These phenomenons indicate the importance of the two inconsistency features. However, how to effectively utilize them has not been revealed comprehensively. Attention-based

Manuscript received December 31, 2020; revised March 29, 2021; accepted May 18, 2021. Date of publication June 7, 2021; date of current version June 29, 2021. This work was supported by the Beijing Natural Science Foundation under Grant JQ18017. This article was recommended for publication by Associate Editor S. Escalera upon evaluation of the reviewers' comments. *(Corresponding author: Gengyun Jia.)*

Gengyun Jia, Xin Ma, and Yuting Xu are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: gengyun.jia@cripac.ia.ac.cn; xin.ma@cripac.ia.ac.cn; yuting.xu@cripac.ia.ac.cn).

Meisong Zheng, Chuanrui Hu, Luoqi Liu, and Yafeng Deng are with 360 AI Institute, Beijing Qihu Keji Company Ltd., Beijing 100020, China (e-mail: zhengmeisong@360.cn; huchuanrui@360.cn; liuluoqi@360.cn; dengyafeng@360.cn).

Ran He is with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: rhe@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TBIOM.2021.3086109

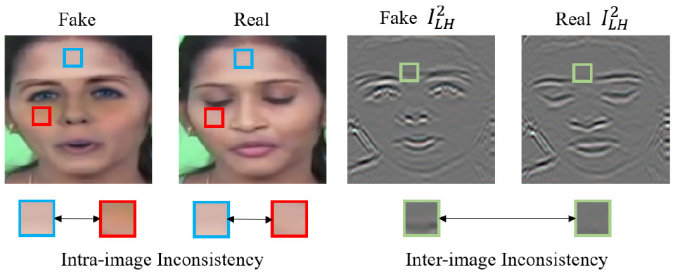

Fig. 1. Two kinds of face forgery clues. The left two are the images in the original RGB color space. We can see that the skin colors in the red and blue rectangles are very different. The right two images are the level-2 stationary wavelet decomposition results. They are column high-pass filtered images. Note that the edge artifacts are captured in the green rectangles.

methods concentrate on the forged regions [23], [24] but may ignore some inter-region relations. To take advantage of the two inconsistency clues, we propose to split the problem into two subtasks, the forgery clue enhancement and extraction, and two methods are designed to achieve them.

We achieve the first task that aims to enhance the forgery clues by investigating frequency domain features. Some previous works pointed out that the forgery clues are more significant in some image frequency sub-band signals [25], [26], [27], such as the high-frequency responses shown in the right two images in Figure 1. Traditional transformations like Fourier Transform (FT) and Discrete Cosine Transform (DCT) represent image frequency spectrum globally regardless of the local frequency information. However, spatial information is important especially in detecting the intra-image inconsistency. Although we can extract spatial features directly from the RGB space, it is hard to bridge the information from the two domains. To this end, we propose to use Stationary Wavelet Decomposition (SWD) to learn space-frequency features. As show in Figure 3, the decomposed images cover different frequency subbands with different spatial resolutions. Furthermore, wavelet based decomposition also extracts multi-direction information. In our model, we use the SWD instead of the traditional Discrete Wavelet Transform (DWT) so that the translation-invariance is maintained. The resolutions of different level wavelet coefficients in the SWD are kept the same as the original images. Therefore, we simply use the coefficients as additional inputs. They are processed by several convolutional layers and finally fused with the RGB image features.

The second task is to effectively extract the two inconsistency features. The inter-image inconsistency comes from the general patterns of each class (real and fake). We use a simple global binary classification task to learn such features. However, to capture the intra-image inconsistency, the model needs to maintain high-resolution features and extract the pixel-level forgery information. To this end, we propose to employ a dual-branch architecture that predicts both image-level forgery labels (classification) and pixel-level forgery labels (segmentation). The classification branch focuses on the inter-image inconsistency and the segmentation branch provides the forgery location information to help extracting intra-image inconsistency. We further employ bilinear pooling to fuse and pool the features from the two branches and finally predict the image-level forgery label. Different from direct feature concatenation or summation, we find that the bilinear pooling is a kind of spatial attentive pooling method. Therefore, it effectively utilize the spatial forgery information learned from the segmentation branch to facilitate the intra-image inconsistency extraction. We conduct sufficient experiments to validate the effectiveness of our proposed methods.

Our contributions in this paper are summarized as follows:

- There are two major features to detect forged faces including the inter-image inconsistency and intra-image inconsistency. We propose to fully utilize them through two tasks that aim to enhance and extract inconsistency features respectively.
- To enhance the inconsistency features, we employ the stationary wavelet decomposed images as additional inputs. The SWD keeps the resolution unchanged and maintain the translation-invariance. It is able to extract localized frequency information comprehensively.
- To extract the inconsistency features, we propose a dual-branch multi-task network to handle the differences between the two features. The two branches learn image-level and pixel-level forgery labels respectively, thus concentrate on different inconsistency features.
- Bilinear pooling is employed to fuse features from the two branches. We find that the bilinear pooling can be regarded as spatial attentive pooling, which takes advantage of the spatial forgery information effectively. With the help of different inconsistency features, the forgery detection performances are finally boosted.

II. Related Work

We briefly review some of the previous works related to our approach in this section.

A. Face Manipulations Detection

By far, the human face is the main biological feature of a person, such as a universal ID card. As a result, there has been a great deal of panic with the advance of artificial intelligence tools that produce realistic facial images that do not exist or modify face attributes in videos in a credible way. DeepFakes is a recent image-forgery technology [21]. Therefore, several orthogonal works have been proposed newly to distinguish between real and manipulated faces [28]. Methods for face forgery detection can be roughly divided into two categories, i.e., discriminative classifiers based methods and data-driven approaches.

The former category usually utilizes diverse semantic discrepancies between the head and the face. Agarwal *et al.* [29] proposed a forensic technique to model facial movements and expression that represent a personal speaking pattern. They build a novel detection model based on the one-class support vector machine (SVM) [30], which can distinguish between real images and manipulated ones. Li *et al.* [31] trained a deep convolutional neural network (DCNN) to find fake face videos based on detection of eye blinking in videos. The authors

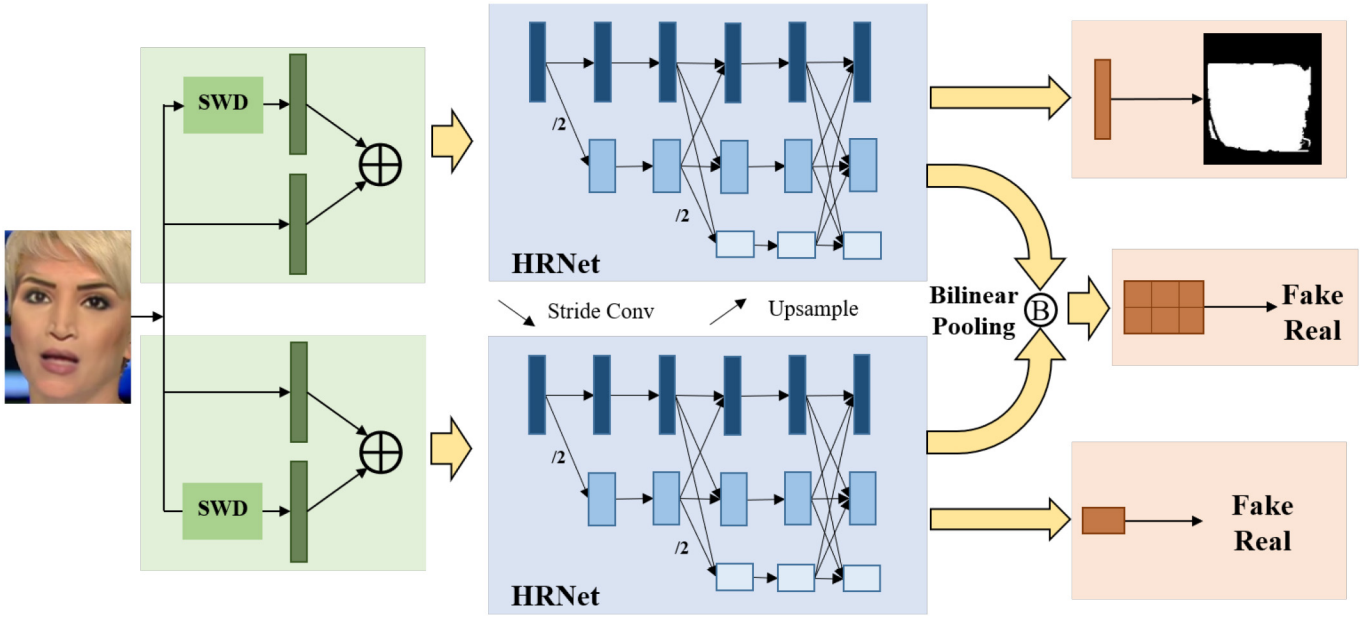

Fig. 2. The pipeline of our proposed method. The face image is fed into the dual-branch networks and each branch has two input routines. One is the normal RGB input, the other decomposes the image into different space-frequency components based on stationary wavelet decomposition. The features of the two input routines are summed up and transmitted to the multi-resolution module, in which the features of different resolutions are extracted (note that Figure is only used to show the structure, it is not the exact number of layers). The output features of different resolutions are processed at each branch to predict pixel-level and image-level forgery labels respectively. The processed features from the two branches are fused with the bilinear pooling to obtain the final image-level label.

believed that this was a physiological signal that didn't show up very well in the synthetic fake video. Later, Li and Lyu [32] observed that there were significant artifacts in DeepFakes videos due to the limited resolution of the generated image used in the warping process. Thus, they directly simulated such artifacts using simple image processing operations and made their method more robust. Matern *et al.* [33] thought that most of computer vision works have limitations when applied to specific, pre-defined scenarios because they would lead to dramatic artifacts in the generated content. Therefore, the authors adopted hand-made visual features to detect image manipulations.

As for data-driven approaches, Rössler *et al.* [18] demonstrated that the XceptionNet-based classifier [34] was superior to all other variants in detecting fakes. Traditional image forensics techniques are often not suitable for video because compression seriously degrades the quality of the data. Thus, Afchar *et al.* [35] presented two networks with few layers to capture the mesoscopic properties of images. Similarly, Zhou *et al.* [36] proposed a two-stream network to detect face tampering. Particularly, GoogLeNet was trained to detect manipulated artifacts in the face classification stream while a patch based triplet network was trained to capture local noise residuals and camera characteristics. Masi *et al.* [28] also proposed a dual-branch structure, namely one branch propagating the original information and the other one amplifying multi-band frequencies. Later, Nguyen *el al.* [37] utilized the multi-task learning strategy to locate and detect manipulated regions simultaneously. Islam *et al.* [38] proposed a dual-order attention model for capturing copy-move location information and exploiting more discriminative features, respectively.

B. Wavelet Transform

In order to perform component analysis on data at multiple scales, wavelet transforms were proposed, which separates data into different space-frequency components. It has been widely used in various computer vision tasks [39]. Nowadays, for different tasks, researchers have proposed various works to combine wavelets with CNNs or other technologies, such as image denoising [40], image super-resolution [15], image style transfer [41] and so on. Deng *et al.* [42] proposed an approach based on wavelet domain style transfer, which achieve perception-distortion trade-off compared with the GAN methods better. Yoo *et al.* [41] proposed a wavelet correction transfer based on whitening and coloring transforms, so that their structural information and statistical characteristics of VGG feature space can be maintained during the stylization process. Besides, Liu *et al.* [40] presented a multi-level wavelet CNN that wavelet transforms are used to reduce the sizes of feature maps. Wavelet transforms also have been used in the image copy-move detection research [43], [44]. The typical framework is to compute the similarities between different local regions in the wavelet domain, and match the regions based on the similarities.

C. Image Semantic Segmentation

Image semantic segmentation, a dense image prediction task, plays an important role in high-level scene understanding. Most of the methods are proposed to extract features of the required spatial resolution and retain the object details. Noh *et al* [45] proposed a novel segmentation approach by utilizing deconvolution to learn powerful representation from

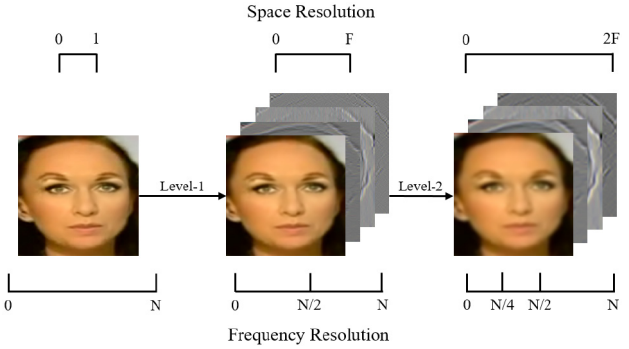

Fig. 3. The SWD with different frequency and space resolutions. In the process, on the one hand, each decomposition output channel only covers half of the input frequency spectrum, indicating a larger frequency resolution. On the other hand, the spatial size of the filter kernel is 2 times larger than the last decomposition level, resulting in a larger space coverage.

low-resolution feature maps. Badrinarayanan *et al.* [46] presented a practical deep fully convolutional neural network consisting of an encoder network, a decoder network followed by a pixel-wise classification layer. Lin *et al.* [47] found that repeated subsampling operations would lead to a dramatic performance decrease. Thus, the author proposed RefineNet, a generic multi-path refinement network, to exploit all the information available explicitly. Meanwhile, other methods focus on aggregating multi-scale contextual information. Liu *et al.* [48] presented a method, named as ParseNet, to add global context to fully convolutional networks by introducing image-level features.

III. Method

In this section, we firstly present the frequency-aware feature extraction based on the stationary wavelet decomposition in Section III-A. Then the dual-branch multi-task learning networks as well as the bilinear pooling strategy to fuse the features are introduced in detail in Section III-B. Finally we introduce other model details including some network architectures and learning processes in Section III-C.

A. Wavelet Based Frequency-Aware Feature Extraction

Many works have analyzed that frequency domain features are helpful to exploit the forgery clues [25]. To extract frequency-aware features, some works use DCT transforms [49]. However, these methods cannot capture the features of different space-frequency resolutions comprehensively. Therefore, we cannot obtain sufficient localized frequency information. To this end, we propose to integrate wavelet-based features in our model.

To be specific, we use stationary 2-D discrete wavelet transform to decompose the input images into different frequency sub-bands. Given the pre-defined wavelet basis, we construct the corresponding 2-D wavelet filters $F^d = \{f^d_{LL}, f^d_{LH}, f^d_{HL}, f^d_{HH}\}$ at the decomposition level d. Take the second filter as an example, the subscript LH means that the filter is a row low-pass and column high-pass filter. Different from classical wavelet transform, the stationary wavelet decomposition upsamples the filters instead of downsampling the filtered images when the

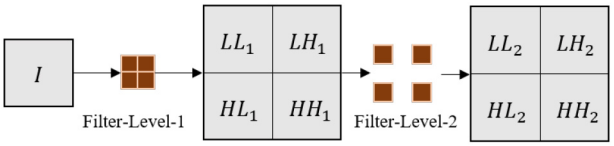

Fig. 4. The SWD can be regarded as applying convolutions with filters of different dilations.

decomposition level $d > 1$, thus the resolution of the decomposed images are kept unchanged. It has two advantages. First, it is convenient to integrate the multi-level SWD outputs into CNNs. Second, the translation-invariance property is maintained. The SWD can be regarded as the dilated convolution with kernel F^d, and the dilation equals to the decomposition level d. Figure 4 gives an example of the filter f_{HH} at different levels. Given an image $I \in \mathbb{R}^{3\times H\times W}$, the first level decomposition generates four subband images $I^1 = \{I^1_{LL}, I^1_{LH}, I^1_{HL}, I^1_{HH}\}$, and the level d outputs I^d are the decomposition results of the I^{d-1}_{LL}. This process is applied to each image channel independently. After each decomposition, the frequency range will be reduced to half of the last level, while the spatial range of each frequency component becomes 2 times larger. Therefore, different decomposition levels reflect the combination of different space-frequency resolutions. This process is visualized in Figure 3.

Once the wavelet decomposed images are obtained, we use convolutional layers to extract frequency-aware features. The two input branches turn the different inputs into features of the same shape. Instead of directly summing up the two features, we employ a channel attention fusion strategy, which calculates the channel attention weights of each branch features. The fused feature is the weighted summation based on the normalized attention weights.

B. Dual-Branch Network With Bilinear Pooling

The two kinds of forgery clues, i.e., the inter-image inconsistency and the intra-image inconsistency, are all helpful in our tasks. However, learning the two clues are significantly different tasks. Intra-image inconsistency needs to capture the different pixel-level patterns, while inter-image inconsistency requires to learn the different responses between real and fake images globally.

Considering the different characters of the two forgery clues, we employ a dual-branch model to learn them respectively as shown in Figure 2. The backbones of the two branches are of the same architecture, but the parameters are not shared. We use the segmentation task to learn the pixel-level forgery labels in the first branch. This branch provides detailed spatial features, which is important to capture the inter-region relations. The second branch directly learns binary classification task to distinguish fake images from real images. Through this architecture, we obtain a high-resolution feature $f_s \in \mathbb{R}^{C_1\times H_1\times W_1}$ from the segmentation branch and a wide feature $f_c \in \mathbb{R}^{C_2\times H_2\times W_2}$ from the classification branch. Obviously, we have $C_1 < C_2$, $H_1 > H_2$ and $W_1 > W_2$, indicating more spatial information in f_s and more semantic information in f_c.

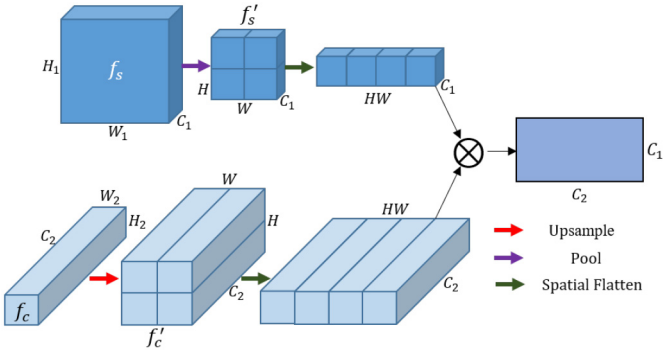

Fig. 5. Bilinear pooling process. The features from the two branches are firstly resized to the same spatial size by interpolation and convolution. Then the spatial dimension is flattened. finally the bilinear pooled feature is obtained by the inner product of the two flattened features.

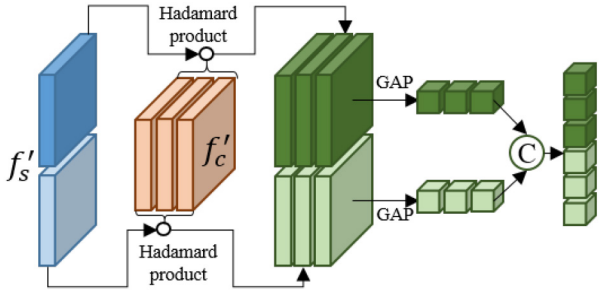

Fig. 6. Another perspective of the bilinear pooling. Suppose there is a two-channel segmentation feature f_s' and a three-channel classification feature f_c'. Each channel of f_s' is multiplied with each channel of f_c' using the Hadamard product. Global average pooling (GAP) is applied to the output. Therefore, it can be regarded as a guided attentive pooling method.

The next question is that how we use the features from the two branches effectively. If there is no need of spatial knowledge, the direct way may be combining global average pooling with feature summation or concatenation. However, the spatial information, especially the local forgery information from the segmentation branch, should not be neglected, otherwise the intra-image inconsistency cannot be effectively utilized. Therefore, we achieve the feature fusion based on bilinear pooling [50].

To conduct the bilinear pooling, the features from the two branches need to be spatially aligned. Therefore, we interpolate the classification features f_c to f_c' and pool the segmentation features f_s to f_s'. f_c' and f_s' are of the same spatial size $H \times W$. We reshape both of the features instead of directly interpolating f_c to reduce the computation burden. The pooling steps are as follows. Firstly, we calculate the outer product of each location feature $f_s'(i) \in \mathbb{R}^{C_1 \times 1}$ and $f_c'(i) \in \mathbb{R}^{C_2 \times 1}$. Then, the global representations are obtained by summing up the results at all positions. The first two steps is written as:

$$f_{bp} = \sum_{i=1}^{HW} f_s'(i) \times f_c'(i)^\top. \tag{1}$$

We achieve the two steps in our case by using the dot product between the spatially flattened features, which is formulated as

$$f_{bp} = T\left(f_s'\right) \cdot T\left(f_c'\right)^\top \tag{2}$$

where $T()$ is the flatten function that turn the features from the shape $\mathbb{R}^{C \times H \times W}$ to the shape $\mathbb{R}^{C \times HW}$. This process is visualized in Figure 5.

Finally, the bilinear pooling outputs are obtained by vectorizing and normalizing the fused outputs f_{bp}

$$f_{bp} = T\left(f_{bp}\right) \tag{3}$$

$$f_{bp} = \text{sign}\left(f_{bp}\right)\sqrt{\left|f_{bp}\right|} \tag{4}$$

$$f_{bp} = \frac{f_{bp}}{\left\|f_{bp}\right\|_2}. \tag{5}$$

To understand why the bilinear pooling can utilize the spatial forgery information, we illustrate its process from another perspective. The bilinear pooling is a kind of attentive pooling that uses multiple spatial attention maps to process the input features. In our model, we regard the segmentation features f_s' as the attention maps and the classification features f_c' as the input features. Unlike using a single forgery mask as the attention map [23], our method introduces richer spatial information to avoid only focusing on the forged regions. Different dimensions of the pooling outputs reflect different combinations of the spatial and semantic responses. As a result, both the intra-image and inter-image inconsistency features are effectively utilized. This perspective is visualized in Figure 6. It is shown clearly that this is a combination of global average pooling and spatial attention.

C. Network Architecture and Model Training

The backbone of our dual-branch model is the High-Resolution Network (HRNet) [51]. We choose this backbone because it learns high-level semantic features and high-resolution spatial features simultaneously. At each network block, the HRNet learns features of different resolutions. Between adjacent blocks, features of different resolutions are summed up to exchange information. We use it to achieve both the classification and the segmentation tasks. In our model, we use the same stem module and high-resolution module as the original HRNet. For the wavelet input branch, we only adjust the number of channels accordingly. We use 3-level wavelet decomposition with Haar wavelet by default. After the features of different resolutions are obtained from the high-resolution modules. We use the network heads for classification and segmentation tasks similar to the original HRNet. As shown in Figure 7, in the segmentation head, the feature maps of different resolutions are interpolated to the same size and concatenated. In the classification head, the feature channels of different resolutions are increased firstly. Then high-resolution features are gradually fused with low-resolution features by convolutions with stride $= 2$. Finally the features of the smallest resolution are globally pooled to predict binary labels. We use the last features before the classification and segmentation layers to perform bilinear pooling.

To learn the pixel-level and image-level labels, we simply use the popular binary cross-entropy loss for classification and

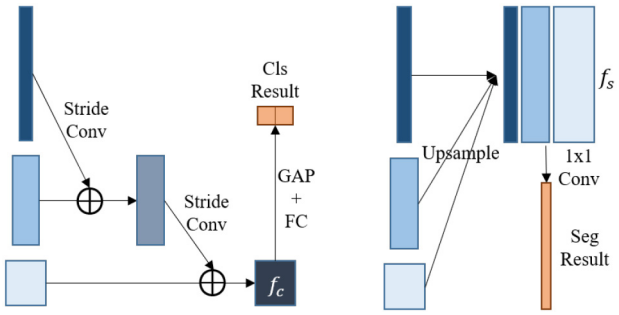

Fig. 7. The network heads for the two branches.The left is the classification head and the right is the segmentation head. The left three features in each head are the outputs of the high-resolution modules.

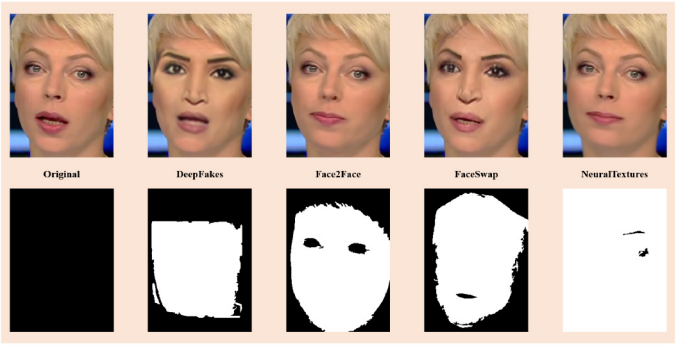

Fig. 8. The ground-truth pixel-level labels of an image and its different forgeries. They are generated through the sliding window SSIM between aligned real and fake faces.

segmentation learning. The loss function is formulated as

$$L = L_{\text{cls}}^{\text{branch2}} + \alpha_1 L_{\text{cls}}^{\text{bilinear}} + \alpha_2 L_{\text{seg}}^{\text{branch1}} \tag{6}$$

where α_1 and α_2 are the trade-off between different loss functions. We simply set them to 1 in our experiments.

IV. Experiments

A. Experimental Settings

1) Dataset: FaceForensics++ [18]. It is a challenging dataset with 1,000 source videos. The number of training videos, test videos and validation videos are 720, 140 and 140 respectively. The videos are collected from YouTube and most of them are the newscast. Each video contains 300-700 frames. There are four kinds of forgery methods manipulating the videos including DeepFakes (DF) [21], FaceSwap (FS) [22], Face2Face (F2F) [55] and NeuralTextures (NT) [56]. The dataset also provides videos of different qualities that are compressed by H.264 algorithm. Following the standard pipeline, we sample 270 frames from each training video and 100 frames from each validating and testing video. When we train the model on all forge types, we only use 100 frames in fake videos to ease the class imbalance problem.

Celeb-DF [20]. It is a high-quality DeepFakes video dataset with 590 real videos and 5, 639 DeepFake videos. The average length of all videos is approximate 13 seconds and the frame rate is 30. 100 frames are randomly sampled in each video for this dataset. Following the standard protocol, we use the designated 518 videos for testing. In the remaining videos, 10% are randomly selected as the validation set.

UADFV [57]. It is a small-scale DeepFakes video dataset containing 49 real videos and 49 fake videos. The average length of these videos is approximately 11.14 seconds. As previous works, we use 35 real and 35 fake videos to train the model, and the remaining videos are left for testing.

2) Evaluation Metrics: Same as the previous works, the classification accuracy (Acc) and the area under the receiver operating characteristic curve (AUC) are used to evaluate models. In the real-world scenarios, we may care about that if we can detect all the fake images in the given data. Therefore, True Positive Rate (TPR) at low False Positive Rate (FPR) is also used. We report the TPR at FPR=0.1. As for the segmentation task, we report the mean Intersection-over-Union (mIoU) results. All the results are the frame-level detection results.

3) Data Pre-Processing: To process the videos and conduct frame-level experiments, we use MTCNN [58] to detect and extract faces from each frame. To avoid incorrectly detected face regions, we use a high threshold 0.99 to filter the detection results. The face region is enlarged by a factor of 1.3 around the center of the detected face. Note that we only detect faces in the real videos, and the detected coordinates can also be used to crop faces in the corresponding fake videos since the face locations and head poses are kept unchanged. But sometimes the video frame resolution may be changed by the forgery algorithms, such as some videos manipulated by the NeuralTextures in the FF++ dataset. Their frame widths are slightly smaller than the corresponding real videos. In this situation, we modify the coordinates to align the face bounding boxes. Based on the paired real and fake face images, we compute the ground-truth pixel-level labels to supervise the segmentation branch. Specifically, we use the structural similarity (SSIM) between paired real and fake images. The SSIM is calculated in a sliding window of size 5×5. The generated SSIM maps are binarized with the threshold 5. Finally we remove the small holes in the maps by the morphology method to increase the smoothness. Figure 8 gives the examples of one cropped face image, the 4 different forged images, and the corresponding pixel-level labels.

There may be some cases in the fake videos that more than one faces exist in some frames and our cropped faces are not the manipulated faces. To avoid the negative influences of such cases, we use the calculated pixel-level labels to correct the image-level labels. Specifically, if the number of value 1 pixel (fake pixel) is smaller than 8, we will regard this face image as a real image. Note that this process aims to deal with the incorrect chosen of multiple faces, not to filter hard samples. In fact, it is shown from Figure 8 that a large portion of pixels are manipulated in true fake images. The number of forged pixels is far beyond the threshold 8.

To train the models with the extracted face images, we resize their long edges to 224 and pad the images to 224×224. Data augmentations are used to improve the generalization ability. Pixel-level augmentations include random noise, blur, brightness, contrast and saturation, and spatial-level augmentations are random cropping, flipping and rotation.

TABLE I
PERFORMANCES OF DIFFERENT VIDEO COMPRESSION LEVELS

Methods	c40			c23			raw		
	Acc	AUC	$TPR_{10\%}$	Acc	AUC	$TPR_{10\%}$	Acc	AUC	$TPR_{10\%}$
Steg.Features [52]	55.98	-	-	70.97	-	-	97.63	-	-
Cozzolino et al. [53]	58.69	-	-	78.45	-	-	98.57	-	-
Bayar and Stamm [54]	66.84	-	-	82.97	-	-	98.74	-	-
MesoNet [35]	70.47	-	-	83.10	-	-	95.23	-	-
DSP-FWA [32]	-	59.15	17.30	-	56.89	14.60	-	-	-
Xception (full) [18]	70.52	-	-	74.78	-	-	82.01	-	-
Xception [18]	81.00	-	-	95.73	-	-	99.26	-	-
TBRN [28]	86.34	86.59	62.48	96.43	98.70	97.95	-	-	-
ours(HRNet-18)	**88.96**	**92.97**	**79.03**	**96.95**	**99.60**	**99.03**	**99.74**	**99.78**	**99.32**

4) Training Details: The HRNet is pre-trained on the ImageNet to accelerate the convergence and improve the generalization ability. The new layers are randomly initialized. We use stochastic gradient descent with initial learning rate 0.01 and momentum 0.9 to train our models. The learning rate is divided by 10 at epoch 10, 13, 15. When the learning rate is changed, we load the previous models with the highest Acc performance in the validation set. The batch size is 128. We use the cross-gpu synchronized batch normalization. The weights of different losses are set to 1.

TABLE II
ACC (%) PERFORMANCES OF DIFFERENT FORGERY METHODS IN C40 COMPRESSION

Methods	DF	F2F	FS	NT
Steg.Features [52]	65.58	57.55	60.58	60.69
Cozzolino et al. [53]	68.26	59.38	62.08	62.42
Rahmouni et al. [59]	80.95	77.30	76.83	72.38
Bayar and Stamm [54]	73.25	62.33	67.08	62.59
MesoNet [35]	89.52	84.44	83.56	75.74
Xception [18]	94.28	91.56	93.7	82.11
ours(HRNet-18)	**98.36**	**94.80**	**97.05**	**84.85**

B. Comparison With Previous Methods

We firstly test our models on different video qualities. In the real-world situation, the videos spread in the social medias are always compressed by popular algorithms such as H.264. Therefore, it is important to test models in different compression levels. The FF++ dataset provides two-level compressed videos (c40 and c23) as well as raw videos. We train our models respectively in each compression level. Table I gives the comparison results with previous methods. We use the directly reported results in the corresponding papers. The Xception (full) method is the Xception trained without cropping the face regions. Our results are based on the backbone networks HRNet-18. All the results are the averaged frame-level performances. We can see from Table that our approach outperforms previous methods in all the evaluation metrics. The performance improvement is significant, especially in the c40 compressed videos. The $TPR_{10\%}$ is improved from previous best 62.48 to 79.03. It is a remarkable improvement, indicating that our model works much better when we require low false positive rate. Note that the classification on the raw video quality is a very simple task that most methods have reached nearly saturated performances. Therefore, the improvement of our model in this situation is limited.

Then we give the results on different forgery types including DF, FS, F2F and NT respectively. The model is trained on all forgery methods at once and evaluated on specific forgery method. As shown in Table II. Our methods achieve better performances in all the four forgery types.

The high performances of our model mainly benefit from the effective enhancement and extraction of different forgery clues. The wavelet-based frequency-aware features enhance the forgery feature representations in the frequency domain. The dual-branch multi-task learning with bilinear fusion effectively captures and fuses the inter-image and intra-image forgery clues.

C. Ablation Study

In this section, we conduct experiments on different variants of our model to prove the effectiveness of the proposed method. The baseline model in our ablation study is the simple one-branch HRNet-18 that predicts the image-level forgery labels without the wavelet-based frequency-aware features. We use AUC and Acc metrics on the c40 compressed videos to show the performances. Besides the experiments on HRNet-18, we add an extra experiment that uses the backbone HRNet-32 to eliminate the effect of the number of parameters. In other words, one-branch HRNet-32 and dual-branch HRNet-18 have a similar number of parameters. We also conduct an experiment that uses the dual-branch architecture but the features are fused through direct summation (features from the two branches are firstly turned in the same width through 1×1 convolution and then globally pooled).

As listed in Table III, the proposed wavelet-based frequency-aware features and the dual-branch network with bilinear feature fusion effectively improve the forgery detection performances. Two important points are observed from Table III. One is that the dual-branch HRNet-18 architectures work better than the single-branch HRNet-32 model. This indicates that the superiority of our model is not only from the increased number of parameters, but also the effective combination of different forgery clues. The other is that

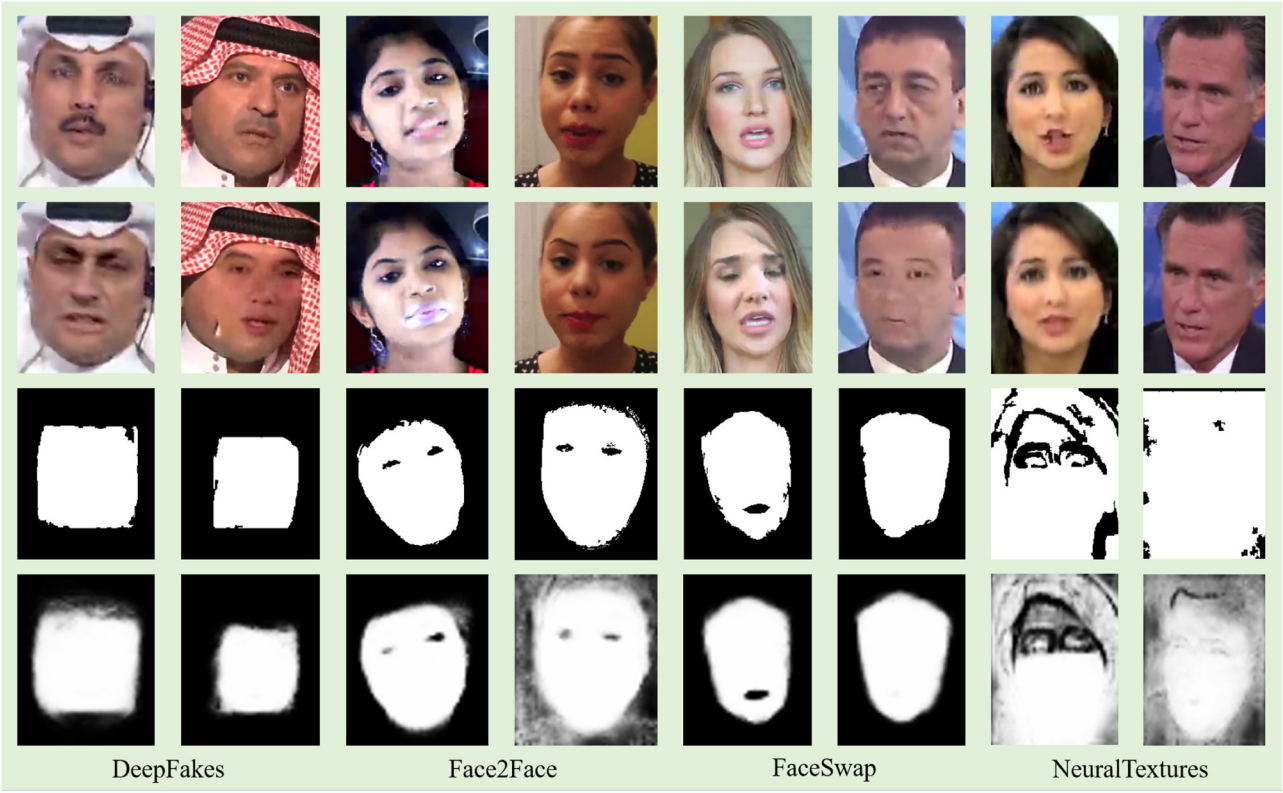

Fig. 9. Some examples of the pixel-level prediction results. They come from the test set of FF++. The model is trained in the c40 compressed level with all four kinds of manipulations. The first row to the fourth row shows the original images, the forged images, the ground-truth pixel-level labels and the predicted soft pixel-level labels respectively.

TABLE III
ABLATION STUDY

HR18	HR32	Wavelet Feature	Dual Branch	Bilinear Fusion	Acc	AUC
✓					86.84	88.96
✓		✓			87.24	90.13
	✓	✓			87.55	91.02
✓		✓	✓		88.35	91.72
✓		✓	✓	✓	88.96	92.97

in the dual-branch models, the bilinear pooling gives higher performances than the direct summation. We think that the direct summation of the two features neglects the spatial relations of the two features, while the bilinear pooling performs attentive pooling that effectively integrates the learned spatial information from the segmentation branch into the classification branch features.

For the wavelet based frequency-aware features, we further conduct experiments of different wavelet basis including Haar and Reverse Biorthogonal wavelets. The results in c40 compressed videos are reported in Table IV. We can see that the different wavelets only cause small performance changes.

TABLE IV
ACC (%) PERFORMANCES OF DIFFERENT WAVELET BASIS

	Haar	Reverse Biorthogonal
Acc	88.96	88.89
AUC	92.97	92.85

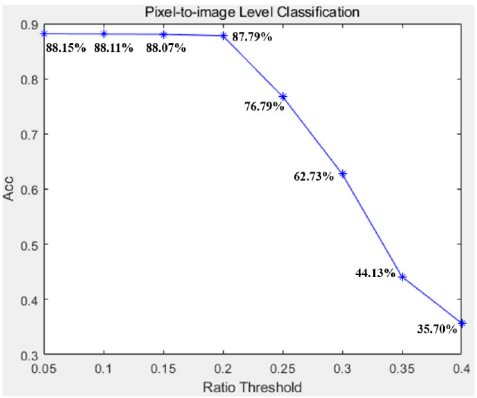

Fig. 10. The Acc performances of pixel-to-image level classification. We use different ratios of the number of forged pixels as the threshold to give the binary results.

D. Pixel-Level Forgery Detection

In our model, the segmentation branch predicts the pixel-level labels. Therefore, we report the segmentation performances qualitatively and quantitatively.

Figure 9 gives the segmentation results on eight randomly selected images. It shows that different forgery types have significantly different forgery regions. The DF algorithm changes the pixel values in a nearly rectangular region that covers

TABLE V
IOU PERFORMACNES (%) OF PIXEL-LEVEL LABEL PREDICTION

	DF	F2F	FS	NT	Real	Avg
c40	87.65	82.41	87.00	46.39	39.81	74.15
c23	92.15	92.90	91.07	58.69	48.14	87.37
raw	95.33	94.42	96.88	60.20	51.66	89.25

the facial features. Therefore, sometimes we can see a clear inconsistency between the forehead and the gills, as shown in Figure 1. The F2F and FS forgery methods manipulate the entire face region, but sometimes the mouth and the eyes are kept unchanged. As for the NT, it changes nearly all pixels in the image. From the last row of Figure, we can see that the model is able to predict the pixel-level labels precisely in most cases. However, the prediction for the NT type is hard. The most possible reason is the irregular small regions in the ground-truth labels.

Table V gives the mean IoU results of different forgery types and video qualities. We train the models in different video qualities respectively. All the forgery types are trained at once but tested separately. The results also show that predicting pixel-level labels for the NT algorithms is hard. This may partly explain the reason why the classification performance improvement on NT in Table II is not significant as the F2F and FS. On the one hand, the segmentation branch could not extract pixel-level label information well enough. On the other hand, the intra-image inconsistency works not well in identifying NT forged images.

Since the pixel-level label has been obtained, Some people may consider using the ratio of the number of forged pixels to get the image-level label. Specifically, we firstly binarize the pixel-level labels to 0 and 1 with the threshold 0.5. Then we compute the ratio of the number of pixels labeled as 1. Finally the ratio is used to classify the images. However, this approach has a major limitation that the ratio threshold is hard to define. Because the manipulated region sizes may be significantly different in different forged images. As a result, we report the classification accuracy on different ratio threshold in Figure 10. The model is trained in c40 quality. As observed in Figure, the Acc performances are good when the ratio threshold is small. The Acc drops with the threshold decreasing. A phenomenon is that when the threshold increases larger than 0.2, the performance degradation becomes really fast. This may tell us that in most of the fake images in the test set, the ratios of the manipulated pixels are larger than 0.2. we further test the AUC to comprehensively evaluate such pixel-to-image level classification. the AUC is only 76.69%, far worse than the image-level results 92.97% reported in Table I. This indicates that this classification method is really unstable.

E. Cross-Domain Test

To validate the generalization performances, we give the cross-domain tests under several situations. In the FF++ dataset, there are four kinds of face forgery methods that are regarded as four different domains. Therefore, we conduct an experiment that the model is only trained with F2F forgery data but tested in all four kinds of forgery images. We show the results in Table VI. The data used in this table are all raw quality, and we use the average AUC performance on the whole validation set to select the best model. It can be seen that our method outperforms the HRNet and Xception models with a huge gap, indicating better generalization ability. Compared with the Face-Xray model, which is specially designed for improving the generalization ability, our method still obtains better performance on DF.

TABLE VI
GENERALIZATION PERFORMANCES OF MODELS TRAINED AND TESTED ON DIFFERENT FORGERY METHODS

Training Set	Models	Test Set AUC (%)			
		DF	F2F	FS	NT
F2F	Xception [18]	87.56	99.53	65.23	65.90
	HRNet [24]	83.64	99.50	56.60	61.26
	Res-Layer1 [60]	84.39	97.66	60.53	79.72
	Face-Xray [24]	98.52	99.06	**72.69**	**91.49**
	Ours	**99.07**	**99.98**	62.98	85.12

TABLE VII
CROSS DATASET GENERALIZATION PERFORMANCES

Models	Training Set	Test Set AUC (%)	
		UADFV	Celeb-DF
Two-steam [36]	Private	85.1	55.7
Meso4 [35]	Private	84.3	53.6
HeadPose [57]	UADFV	89.0	54.8
FWA [32]	UADFV	97.4	53.8
Multi-task [37]	FF	65.8	36.5
Xception [23]	FF++	80.4	38.7
	DFFD	75.6	63.9
	UADFV	96.8	52.2
	UADFV+DFFD	97.5	67.6
Xception+Reg [23]	DFFD	84.2	64.4
	UADFV	98.4	57.1
	UADFV+DFFD	98.4	71.2
Ours	FF++(c23)	80.8	**72.3**
	FF++(c40)	75.6	70.2
	UADFV	**99.9**	53.9
	Celeb-DF	**99.6**	**99.5**

We also test our models on the UADFV and Celeb-DF datasets with different training sets. The results are given in Table VII. Four different training sets are used to train our model, including FF++ with c23 and c40 qualities, UADFV and Celeb-DF. For the AUC metric on the UADFV test set, we can see that our model obtain the best performances in both in-dataset test (99.9%) and cross-dataset test (99.6%). Surprisingly, the model only trained in Celeb-DF performs well in the UADFV test set. One possible reason is that the domain gap is not significant. As for the Celeb-DF dataset, our model trained with FF++ (c23) outperforms previous methods in the cross-dataset condition. These experimental results demonstrate the superiority of our model's generalization ability.

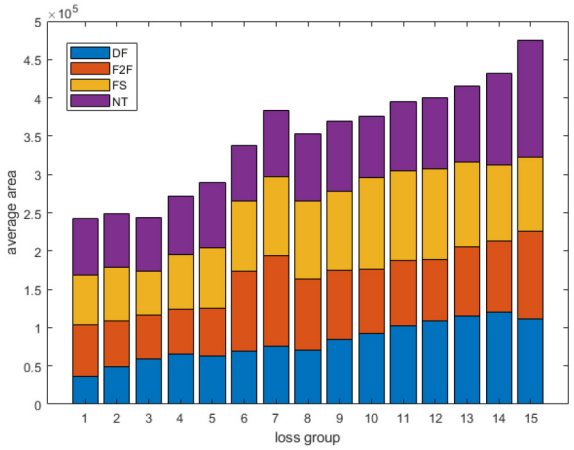

Fig. 11. The relations between the testing losses and the image resolutions. The images are ranked in the testing loss descending order and divided into adjacent groups. For example, the losses in the group 1 are all larger than the losses in the group 2. The y axis is the average product of the image width and height (the image area). It is shown that the larger losses tend to come from smaller images.

F. Failure Cases Analysis

When applying the face forgery detection model to the real-world condition, it is important to analyze the factors that may lead to failed predictions. Obviously, we can see from the results on FF++ dataset that the forgery method and video quality are the most important factors. The video quality deterioration significantly improves the detection difficulty. As listed in Table I, when using the raw videos, nearly all the methods reach accuracy beyond 0.95. But under the compression level c40, the accuracy of all the models in Table I is below 0.90. As for the different forgery methods, Table II clearly shows the performance differences. It can be seen that NT is really hard to detect.

Besides these factors that has been discussed in some previous works, we also notice that the cropped face image resolution may be related to the detection performance. For each forgery method, we record the loss of each test image and sort the images in the loss descending order. Then we divide the images into 28 consecutive groups according to the sorted order and each group contains 500 images. Finally we show the average image area (pixel number) of the first 15 groups in Figure 11. Figure shows the roughly negative correlation between the loss and the image resolution. In other words, smaller image resolution may tend to get higher loss in the model. Therefore, face resolution is an important factor in face forgery detection.

V. Conclusion and Future Works

We present a method for face forgery detection based on the stationary wavelet decomposition and the dual-branch network with the bilinear feature fusion. We analyze that the face forgery detection is based on intra-image and inter-image inconsistency, and propose to use two tasks, feature enhancing and extraction, to better utilize them. By introducing the stationary wavelet decompsition, our model extracts features of different space-frequency resolutions, thus enhances the inconsistency features. We design a dual-branch multi-task learning network to learn both pixel-level and image-level forgery labels. Therefore, the features reflecting the two inconsistencies are effectively extracted. Finally, we employ the bilinear pooling to effectively combine the features from the two branches to classify the images. Our model gives the results that outperform previous state-of-the-art methods with various evaluation metrics. In the future, we will continue our works and find more effective methods to make use of the intra-image and inter-image inconsistencies.

References

[1] H. Fu and X. Cao, "Forgery authentication in extreme wide-angle lens using distortion CUE and fake saliency map," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 4, pp. 1301–1314, Aug. 2012.

[2] H. Cao and A. C. Kot, "Accurate detection of demosaicing regularity for digital image forensics," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 4, pp. 899–910, Dec. 2009.

[3] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of CFA artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1566–1577, Oct. 2012.

[4] S. Lyu, X. Pan, and X. Zhang, "Exposing region splicing forgeries with blind local noise estimation," *Int. J. Comput. Vis.*, vol. 110, no. 2, pp. 202–221, 2014.

[5] M. Kobayashi, T. Okabe, and Y. Sato, "Detecting forgery from static-scene video based on inconsistency in noise level functions," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 4, pp. 883–892, Dec. 2010.

[6] Z. Fan and R. L. De Queiroz, "Identification of bitmap compression history: JPEG detection and quantizer estimation," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 230–235, Feb. 2003.

[7] H. Farid, "Exposing digital forgeries from JPEG ghosts," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 1, pp. 154–160, Mar. 2009.

[8] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1053–1061.

[9] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–6.

[11] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, "High fidelity face manipulation with extreme poses and expressions," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2218–2231, Jan. 2021.

[12] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "Dual variational generation for low shot heterogeneous face recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 2670–2679.

[13] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "DVG-Face: Dual variational generation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 18, 2021, doi: 10.1109/TPAMI.2021.3052549.

[14] X. Ma, X. Zhou, H. Huang, Z. Chai, X. Wei, and R. He, "Free-form image inpainting via contrastive attention network," in *Proc. Int. Conf. Pattern Recognit.*, 2020, pp. 1–9.

[15] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet domain generative adversarial network for multi-scale face hallucination," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 763–784, 2019.

[16] Y. Li, H. Huang, J. Cao, R. He, and T. Tan, "Disentangled representation learning of makeup portraits in the wild," *Int. J. Comput. Vis.*, vol. 128, no. 8, pp. 2166–2184, 2020.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[18] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1–11.

[19] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 2886–2895.

[20] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3207–3216.

[21] *DeepFakes*. Accessed: Dec. 31, 2020. [Online]. Available: https://github.com/deepfakes/faceswap
[22] *FaceSwap*. Accessed: Dec. 31, 2020. [Online]. Available: https://github.com/MarekKowalski/FaceSwap/
[23] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5781–5790.
[24] L. Li *et al.*, "Face *x*-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5001–5010.
[25] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3247–3258.
[26] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7890–7899.
[27] T. Dzanic, K. Shah, and F. Witherden, "Fourier spectrum discrepancies in deep network generated images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, p. 33.
[28] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 667–684.
[29] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 38–45.
[30] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.
[31] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," 2018. [Online]. Available: arXiv:1806.02877.
[32] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 46–52.
[33] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, 2019, pp. 83–92.
[34] F. Chollet, "XCeption: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.
[35] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018, pp. 1–7.
[36] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1831–1839.
[37] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *Proc. IEEE 10th Int. Conf. Biometric Theory Appl. Syst. (BTAS)*, 2019, pp. 1–8.
[38] A. Islam, C. Long, A. Basharat, and A. Hoogs, "DOA-GAN: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4676–4685.
[39] J. Z. Wang, "Wavelets and imaging informatics: A review of the literature," *J. Biomed. Informat.*, vol. 34, no. 2, p. 129, 2001.
[40] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-CNN for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 773–782.
[41] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9036–9045.
[42] X. Deng, R. Yang, M. Xu, and P. L. Dragotti, "Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3076–3085.
[43] G. Muhammad, M. Hussain, K. Khawaji, and G. Bebis, "Blind copy move image forgery detection using dyadic undecimated wavelet transform," in *Proc. 17th Int. Conf. Digit. Signal Process. (DSP)*, 2011, pp. 1–6.
[44] Y. Wang, K. Gurule, J. Wise, and J. Zheng, "Wavelet based region duplication forgery detection," in *Proc. 9th Int. Conf. Inf. Technol. New Gener.*, 2012, pp. 30–35.
[45] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.
[46] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[47] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1925–1934.
[48] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015. [Online]. Available: arXiv:1506.04579.
[49] K. Xu *et al.*, "Learning in the frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1740–1749.
[50] T.-Y. Lin, A. R. Chowdhury, and S. Maji, "Bilinear convolutional neural networks for fine-grained visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1309–1322, Jul. 2018.
[51] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.
[52] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, Jun. 2012.
[53] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection," in *Proc. 5th ACM Workshop Inf. Hiding Multimedia Security*, 2017, pp. 159–164.
[54] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. 4th ACM Workshop Inf. Hiding Multimedia Security*, 2016, pp. 5–10.
[55] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2387–2395.
[56] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, 2019.
[57] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 8261–8265.
[58] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[59] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *Proc. IEEE Workshop Inf. Forensics Security (WIFS)*, 2017, pp. 1–6.
[60] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 103–120.

Gengyun Jia received the B.E. degree in communication engineering from Shandong University, Jinan, China, in 2015, and the M.S. degree in information and communication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2018. He is currently pursuing the Ph.D. degree in computer application technology with the University of Chinese Academy of Sciences, Beijing, and with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include image understanding, media forensics, and machine learning.

Meisong Zheng received the B.E. degree from Central South University, Changsha, China, in 2009, and the M.S. degree from Shandong University, Jinan, China, 2012, and the Ph.D. degree in Computer Science from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2016. She is currently working with AI Institute, Qihoo 360 Technology, Beijing. Her research interests include multimedia, computer vision, and deep learning.

Chuanrui Hu received the M.S. degree from the School of Electrical Engineering and Automation, Anhui University in 2019. He is currently the Director of Multimedia Content Audit Group, Qihoo 360 Corporation. His research interest mainly in computer vision and deep learning.

Xin Ma received the B.E. degree in electronic information engineering from Jiangsu University, Jiangsu, China, in 2018. He is currently pursuing the M.S. degree in computer technology with the University of Chinese Academy of Sciences, Beijing, China, and with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include image super-resolution, image inpainting, and machine learning.

Yuting Xu received the B.E. degree in software engineering from the Taiyuan University of Technology, Shanxi, China, in 2018. She is currently pursuing the M.S. degree in electronic and information with the University of Chinese Academy of Sciences, Beijing, China, and with the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include media forensics and machine learning.

Luoqi Liu received the Ph.D. degree from the Department of Electrical and Computer Engineering, National University of Singapore in 2015. He is currently the Director of CV Group, Qihoo 360 Corporation. His research interests are mainly in computer vision, multimedia, and deep learning.

Yafeng Deng is currently the Head of the AI Research Institute and Search Business Division, Qihoo 360 Technology Company Ltd. He is an Alumnus of Tsinghua University and possesses over 18 years of research experience in the field of computer vision and artificial intelligence. He has published 11 papers and has applied for 135 patents in China, of which 98 have been approved. He was the CTO with DeepGlint Information Technologies Company Ltd., and also held positions as an Research and Development Specialist with multiple different companies, including Baidu Deep Learning Research Institute, Alibaba Cloud Computing, and Shanda Innovation Institute. He is responsible for business relating to facial recognition, video content analysis, action recognition, image search, speech recognition, robotics, and deep learning with applications in natural language processing and recommendation systems. He has lead teams to first place positions in multiple international and domestic mainstream competitions, including FRVT, LFW, FDDB, KITTI, and PRCV2019. He is a member on the CCF-Computer Vision Technical Committee, as well as the Computer Vision Standards Body. He also held the position of Smart Safety (Machine Vision) Specialist in Huawei Technologies Company Ltd., and was also a member of the Intel Advisory Committee.

Ran He (Senior Member, IEEE) received the B.E. and M.S. degrees in computer science from the Dalian University of Technology, Dalian, China, 2001 and 2004, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2009. Since September 2010, he has joined NLPR where he is currently a Full Professor. His research interests focus on information theoretic learning, pattern recognition, and computer vision. He serves as an Associate Editor of the *Neurocomputing* (Elsevier), and serves on the program committee of several conferences.