IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 August 2023, accepted 9 October 2023, date of publication 13 October 2023, date of current version 24 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324541

RESEARCH ARTICLE

IRL-Net: Inpainted Region Localization Network via Spatial Attention

AMIR ETEFAGHI DARYANI[1], (Student Member, IEEE), MAHDIEH MIRMAHDI[2], (Member, IEEE), AHMAD HASSANPOUR[3], (Student Member, IEEE), HATEF OTROSHI SHAHREZA[4,5], (Graduate Student Member, IEEE), BIAN YANG[3], (Member, IEEE), AND JULIAN FIERREZ[6], (Member, IEEE)

[1]Department of Electrical Engineering, Amirkabir University of Technology, Tehran 15875-4413, Iran
[2]Faculty of Computer Engineering, University of Isfahan, Isfahan 81746-73461, Iran
[3]Department of Information Security and Communication Technology, Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway
[4]Biometrics Security and Privacy Group, Idiap Research Institute, 1920 Martigny, Switzerland
[5]School of Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
[6]School of Engineering, Universidad Autónoma de Madrid, 28049 Madrid, Spain

Corresponding author: Ahmad Hassanpour (Ahmad.Hassanpour@ntnu.no)

This work was supported by the Project Privacy Matters (PRIMA) under Grant H2020-MSCA-ITN-2019-860315. The work of Hatef Otroshi Shahreza was supported by the H2020 TReSPAsS-ETN Marie Skłodowska-Curie Early Training Network under Grant 860813. The work of Julian Fierrez was supported by the Project Biometrics and Behavior for Unbiased and Trusted AI with Applications (BBforTAI) under Grant PID2021-127641OB-I00 MICINN/FEDER.

ABSTRACT Identifying manipulated regions in images is a challenging task due to the existence of very accurate image inpainting techniques leaving almost unnoticeable traces in tampered regions. These image inpainting methods can be used for multiple purposes (e.g., removing objects, reconstructing corrupted areas, eliminating various types of distortion, etc.) makes creating forensic detectors for image manipulation an extremely difficult and time-consuming procedure. The aim of this paper is to localize the tampered regions manipulated by image inpainting methods. To do this, we propose a novel CNN-based deep learning model called IRL-Net which includes three main modules: Enhancement, Encoder, and Decoder modules. To evaluate our method, three image inpainting methods have been used to reconstruct the missed regions in two face and scene image datasets. We perform both qualitative and quantitative evaluations on the generated datasets. Experimental results demonstrate that our method outperforms previous learning-based manipulated region detection methods and generates realistic and semantically plausible images. We also provide the implementation of the proposed approach to support reproducible research via https://github.com/amiretefaghi/IRL-Net.

INDEX TERMS Image forensics, image inpainting, image manipulation detection.

I. INTRODUCTION

Image manipulation has become very convenient and ubiquitous nowadays due to the availability of some easy-to-use tools like Adobe Photoshop. Some image manipulation techniques can lead to misinterpretation, and thus malicious use of the visual content, e.g.: moving some elements from one region to another region (copy-move) [1], [2], [3], copying elements from one image and pasting them on another image (splicing) [4], [5], [6], [7], [50], and removal of unwanted elements [8]. Some of these techniques leave very few traces behind, making the detection of manipulated regions very challenging. For instance, recent learning-based inpainting methods attempt to semantically fill the corrupted region based on the overall scene, and the missed region is continuously structured with uncorrupted regions. Particularly, when the aim is to inpaint small missing regions, the outputs of these methods visually look very realistic. Even recent advances in inpainting methods show that they can fill large missing areas with meaningful

The associate editor coordinating the review of this manuscript and approving it for publication was Davide Patti.

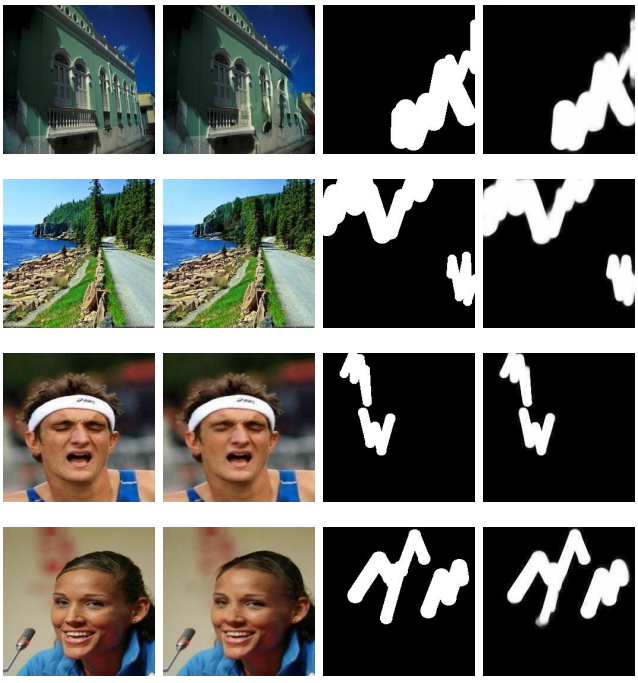


FIGURE 1. Examples of the proposed IRL-Net method predicting manipulated regions. The examples in the first two rows are from Places2 [10] and the other two from CelebA [11] datasets which are inpainted by the GC method [12] using the indicated Mask (c) , and then detected with our method (d).

structures and objects that do not exist anywhere else in the image [9].

Such advancements make the manipulation detection a very challenging process [13], especially when the aim is not only to discriminate manipulated images from the authentic ones, but also to pinpoint tampered regions at the pixel level [14]. Notably, different categories of GAN-based inpainting methods [15] are trained using various sizes of masks which enable them to predict small or large masked regions, leading to, as shown in Fig. 1, inpainted images visually plausible with almost no manipulation traces left around or inside the inpainted regions [16]. In this paper, we address tampering localization by focusing on unveiling the tampering traces left by inpainting methods.

Due to the widespread usage of deep learning in many fields, the community of multimedia forensics has been inspired and driven to investigate if it is feasible to compel a convolutional neural network (CNN) to learn manipulation detection characteristics and record pixel value dependencies caused by image tampering operations. The existing deep learning-based methods can be classified into two main categories. First, the methods that benefit from a noise map of the input image generated either by pre-defined high-pass filters [17] or trainable counterparts [18], [19]. For instance, the spatial-domain rich model (SRM) [17], a non-trainable layer, has been proposed to capture the local dependency changes caused by manipulation techniques. The output of such layers is fed into a deep neural network (DNN), either alone [19], [20] or together with the input RGB image [21], [22], [23]. The approaches within the second category usually do not use any special layer for detecting noise from the input image. Instead, they utilize different network structures like fully convolutional networks (FCN) [24], faster R-CNN [25], and long short-term memory (LSTM) cells [26].

In this paper, we propose a novel method, named IRL-Net, which uses the former approach where high-level features are extracted from both the RGB image and a high-pass filtered version of the RGB image. The extracted features from RGB and high-pass filter can be concatenated together at different stages for further processing. The concatenate stage of the two feature branches is understudied and can be categorized into three classes called early, middle, and late fusion [27]. In this paper, we will investigate different fusion strategies in that regard, and our experiments show that the early fusion model performs better than the other two fusion types.

We also perform end-to-end training to learn the most discriminative features between manipulated and non-manipulated regions through back-propagation using ground truth labels and image mask information. To improve the performance of detecting and localizing manipulated image regions, different kinds of CNN-based approaches have been presented to classify the image patch and pixel-wise segmentation, and different inputs to the network are taken into consideration. The proposed model achieves promising results in patch classification, as well as in localizing manipulated regions at pixel level. Our main contributions can be summarized as:

- In this work, we propose a novel effective end-to-end solution for localizing the manipulated regions generated by inpainting methods. Our DNN-based model, called IRL-Net, benefits from the advantages of a new proposed attention layer. The code for our proposed method is available on GitHub.[1]
- We utilize two effective blocks called attention and up-scaling to predict very high-quality outputs. The attention block is responsible to extract more informative features and the upscaling block placed in the Decoder module assists to generate a super-resolution with minimum checkerboard artifact issues in the output.
- The required datasets for training and validation have been generated using two publicly available datasets called Places2 and CelebA. Moreover, to inpaint the masked regions, three well-known and recently proposed inpainting methods have been used.

II. RELATED WORKS

In this section, we first review recent works developed for image inpainting and then present the methods that concern the localization of inpainted areas.

[1] https://github.com/amiretefaghi/IRL-Net

A. IMAGE INPAINTING

The remarkable progress of DNNs provides the image inpainting task with a great opportunity to produce very realistic results, making it very arduous for human eyes to recognize inpainted regions. These remarkable results have made image inpainting to grow significantly in specific application areas, e.g., face inpainting [28], and scene inpainting [29]. Nowadays, image inpainting can assist in removing, restoring, or reconstructing lost or corrupted part of the image. Generally, existing inpainting models differ in terms of the network structure. For example, some methods follow the coarse-to-fine technique [12], [30] to gradually refine the generated images. The two consecutive stages (i.e., coarse and fine) respectively learn the missing regions at the coarse stage and further refine the whole image at the fine stage. Besides the coarse-to-fine structure, another well-known structure called coarse-and-fine has the aim of extracting global semantic information as well as multi-level local features in parallel [30], [31].

B. LOCALIZING TAMPERED REGIONS

Detecting manipulated regions is a binary classification problem where the classifier should decide about each pixel: tampered or not. Traditional DNN-based solutions have tried to localize the pixels manipulated by inpainting methods usually had poor performance, mainly because they used the specific content of the image at hand as their main information source instead of content-independent features.

More recently, some approaches have been proposed to look for the footprint of tampered pixels in a residual space not focusing on the specific content of the image but interpreting that the tampered regions mostly differ from the untouched parts in terms of their noise distribution. In order to construct a noise map, methods are categorized into two groups: non-trainable and trainable. In [19] the noise map provided by high pass filters (pre-filtering) is fed to four Residual blocks followed by upsampling modules to achieve pixel-wise prediction. However, Bayar et al. [18] proposed a constrained convolutional layer (called Bayar layer) that adaptively learns to suppress the image's content and learns manipulation detection features. Several methods [21], [22], [25], [44], [45], [46] were proposed to leverage both the noise map and content of the image to reduce the risk of losing other useful information in the original RGB view. Zhou et al. [25] proposed a two-stream fast R-CNN for image manipulation detection, the RGB image, and its noise counterpart generated by the spatial rich model (SRM) [17]. One stream extracts features to find tampering artifacts, and the other one discovers noise discrepancies between the tampered region and untouched parts. The Manipulation Tracing Network (ManTra-Net) [22] uses not only the RGB view but also two noise counterparts: SRM [17] and Bayar layer [18]. ManTra-Net decomposes to a feature extraction part followed by a LSTM based detection module. The Spatial Pyramid Attention Network (SPAN) [21], similar to ManTra-Net [22], leveraged SRM [17], Bayar layer [18], and RGB view for its feature extraction module's fed. The Image Inpainting Detection Network (IID-Net) [8] leveraged the incorporation of the SRM layer, Pre-Filtering layer [19], and Bayar layer [18]. The fusion of three streams was fed to an extraction block designed by the Neural Architecture Search (NAS) algorithm, followed by a decision block encompassing global and local attention modules to reduce intra-class inconsistency.

On another front, edge-supervised approaches have been recently presented in some related papers [23], [24], [32], [47], [48], [49] aiming to trace various manipulation types. Methods that exploit edge-supervised techniques look for boundaries around tampered areas. Nevertheless, this strategy is not practical to fulfill the purpose of localizing high-quality inpainting methods whose boundaries are almost imperceptible. For instance, Multi-View Multi-Scale Supervised Networks (MVSS-Net) [23] exploits tampering boundary artifacts by using an edge-supervised method alongside the noise view of the input image and RGB view.

The methods combining the RGB stream and its noise counterpart have a fusion part categorized into early, middle, or late-stage fusion [27]. Accordingly, ManTra-Net [22], SPAN [21], and IID-Net [8] have an early fusion part for concatenating features of the two corresponding streams. However, other methods [23], [25] proposed late fusion so each stream provides deeper-layer features before concatenation. Notably, to the best of our knowledge, middle fusion has not been studied for inpainting manipulation localization.

Due to the limited availability of implementations for other methods, we conduct a comparative analysis between the results of our proposed method and two recent alternatives, namely MVSS-Net [23] and LDICN [19].

III. PROPOSED METHOD

The overview of our proposed method, including three main modules named Enhancement, Encoder, and Decoder, is shown in Fig. 2. The Enhancement module receives the manipulated image and tries to enhance inpainting traces. Then, the Encoder module, which is using a residual block, is intended to extract high-level features that assist to discriminate the manipulated region from the rest of the image. Finally, the Decoder module generates a predicted inpainting mask with the help of an Attention block and Pixel-Shuffle upscaling blocks.

A. ENHANCEMENT MODULE

Generally, standard convolution layers learn features to represent the contents of input images rather than extracting the required features for detecting the traces left behind by inpainting methods. Notably, the majority of these traces are hidden in local noise distributions, and usually RGB channels are not sufficient to deal with all types of manipulation traces. Considering this, with the aim of suppressing ineffective content of the input image and more significantly capturing the inpainting traces, a special predesigned layer called

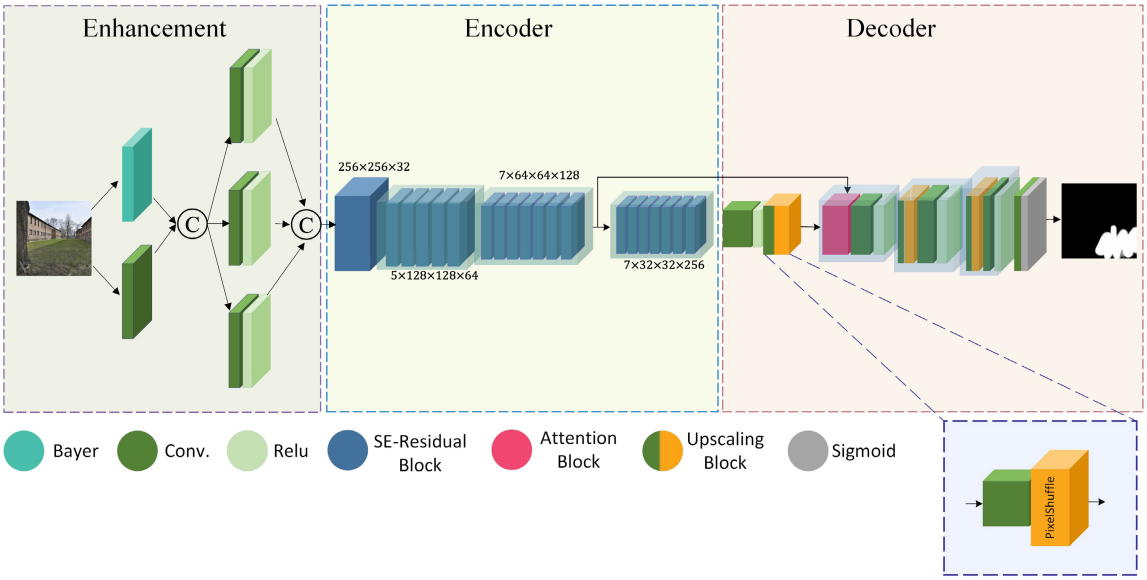


FIGURE 2. Overview of our proposed inpainted region localization Network (IRL-Net) architecture including three main modules called Enhancement, Encoder, and Decoder.

Bayar [18] has been used. Therefore, we adopt the Bayar layer as one of the early layers of the Enhancement module to learn low-level prediction residual features to detect the inpainting traces. The Bayar layer reaches this goal by adding specific constraints to the standard convolution layer in the following way. Let W_b^i represent the ith channel (for RGB input image $i = 1, 2, 3$) channel of the weights W_b in the Bayar layer. The following the constraints are enforced on each channel of W_b before each training iteration:

$$\begin{cases} W_b^i(0,0) = 1 \\ \sum_{m,n\neq 0} W_b^i(m,n) = 1 \end{cases} \qquad i = 1, 2, 3 \qquad (1)$$

where $W_b^i(0,0)$ indicates the center of ith channel of the weights W_b in the Bayar layer. Then, we concatenate the extracted features by the Bayar and convolution layers and feed them to a Multi-Scale Convolution (MSC) block. As the scale of generated noise (manipulated traces) by inpainting methods vary, introducing multi-scale feature extraction can help to learn more robust convolutional filters, and thus more informative features will be extracted. The MSC block has three convolution layers with the size of $x \times 3 \times 3$, $x \times 5 \times 5$, and $x \times 7 \times 7$, where x is the number of filters in each layer. Finally, we concatenate the output of each layer and transmit it to the next module (i.e., the Encoder module).

B. ENCODER MODULE

To extract high-level features, an Encoder module, including four residual units each of which is filled with residual layers, has been placed after the Enhancement module. Using the residual architecture [33] to avoid vanishing/exploding gradients, the Encoder module can assist to extract more abstract features. As shown in Fig. 3 (a), a residual block

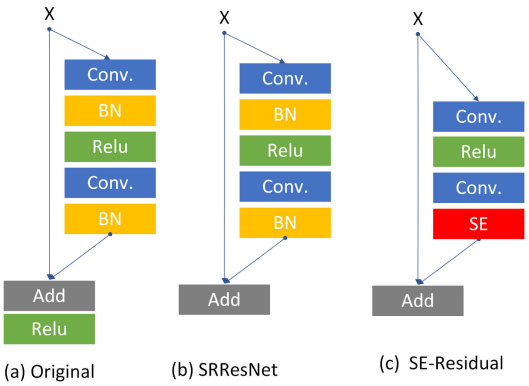


FIGURE 3. Different version of Residual blocks a) The original version of Residual block b) The last Relu block has been removed c) Batch Normalization block removed and SE Block added.

is a stack of layers set in a way that the output of a layer is taken and added to another layer deeper in the block. The non-linearity is then applied after adding it together with the output of the corresponding layer in the main path. This bypass connection is known as shortcut or the skip connection.

1) SE-RESIDUAL BLOCK

An effective residual block called SE-Residual proposed by [34] has been used to prevent: a) decreasing the flexibility of the network for extracting features, and b) increasing the number of feature maps leading to numerically unsuitability during the training phase. Here, the SE-Residual block has been used besides the two original [33] and SRResNet [35] residual blocks. In the SE-Residual block, the

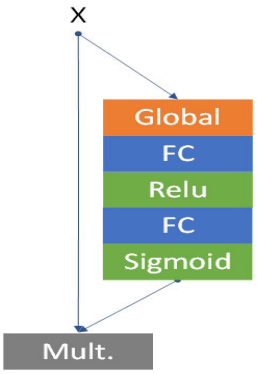


FIGURE 4. Squeeze and excitation block.

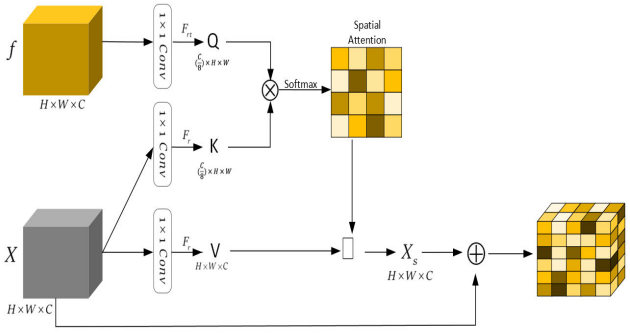


FIGURE 5. Attention block.

batch normalization layers are removed to provide various flexibility ranges for extracting features, as Nah et al. [36]. Moreover, instead of increasing the number of convolution layers or feature maps to improve the performance, this SE-Residual unit improves the representational power of the network by enabling it to perform dynamic channel-wise feature recalibration. As shown in Fig. 4, as part of this process, the block squeezes each input channel into a single numerical value using Global Average Pooling. The second step of this block is to extract information from the input by two Fully-Connected layers (FC). The first FC layer exploits the ReLU activation function and reduces the output channel complexity. A sigmoid activation function is used in the second FC layer, which gives each channel a smooth gating function. In the end, the block weights each feature map of the input according to its channels: the ''excitation''.

C. DECODER MODULE

To map the learned high-level features extracted by the Encoder module into low-level discriminative information, the Decoder module has been placed as the last module of our proposed method. The output of the Decoder module is a mask image (black and white image) showing the manipulated region by white pixels (positive class) and pristine regions by black pixels (negative class). The Decoder module receives high-level features at lower-scale in comparison to low-level mask images, therefore it should upscale features to generate an appropriate mask image. This upscaling process is performed by a PixelShuffle block (described in the next section) as shown in Fig. 2. During this process, misclassified pixels may be generated in the mask image, due to the ineffectiveness of convolutional neural networks in modeling long-term feature correlations. To track this problem, many attention blocks have been proposed and used recently in the decision phase of networks. In this line of work, we designed an attention block and use it in the decoder to generate the mask image in an accurate way. This attention block aims to reduce the number of misclassified pixels through a very effective technique: using knowledge of the Encoder module to assist the Decoder module to build an appropriate features map. The attention map is directly computed on the decoder and encoder features. After obtaining the attention scores, we use these to compute attention on decoder features, as shown in Fig. 5.

1) UPSCALING BLOCK

A special upscaling block called PixelShuffle has been used to construct a high-accurate output (mask image). The traditional upscaling usually starts with some kind of interpolation (e.g., bilinear) which usually leads to checkerboard artifact issues. To reduce those artifacts Shi et al. [37] introduced PixelShuffle which is an operation used in super-resolution models to implement efficient sub-pixel convolutions with a stride of $1/r$. Specifically, PixelShuffle rearranges elements in a tensor of shape (None, $W, H, C \times r^2$) to a tensor of shape (None, $W \times r, H \times r, C$). As shown in Fig. 2, we have used one PixelShuffle operation for $\times 2$ upscaling and two PixelShuffle operations for $\times 4$ upscaling.

2) ATTENTION BLOCK

This attention block is inspired by self-attention, hence, this attention block has three variables known as Query/Key/Value. Queries are a set of vectors you want to calculate attention for. Keys are a set of vectors you want to calculate attention against. Dot product multiplication gives you a set of weights (also vectors) indicating how attended each query is against Keys. Based on our purpose and this definition, we use features of the Encoder module as query and futures of the Decoder module as key. Under this condition, the attention block uses the knowledge of the Encoder module to assist the Decoder module to build an appropriate features map. As shown in Fig. 5, the features of the Decoder module (X) and Encoder module (f) are fed into 1×1 convolution layers, and the outputs are then reshaped to the feature maps Q and K, respectively:

$$Q = F_{rt}(W_q f) \tag{2}$$

$$K = F_{rt}(W_k X) \tag{3}$$

where F_r is the reshape function to convert the height and width dimensions of the feature map into one dimension while F_{rt} is the F_r followed by a transpose operation. Then,

the spatial attention of each location is defined as follows:

$$A^{i,j} = \frac{exp(K_i Q_j)}{\sum_{i=1}^{N} exp(K_i Q_j)} \tag{4}$$

where each element $A_s^{i,j}$ of the spatial attention map A_s represents the correlation extent between the ith position and the jth position to model the long-range dependency. Simultaneously, X is also fed into a 1×1 convolution layer to generate a feature map, which is then reshaped. A matrix multiplication operation between V and A_s is performed and then the result is reshaped back to $R^{W \times H \times C}$ for generating the attended feature map X_s as follows:

$$X_s = VA_s \tag{5}$$

Finally, we multiply X_s by a scale factor α and add with the input feature map X to generate the final output of the attention block as follows:

$$O = \alpha X_s + X \tag{6}$$

where α is a learnable parameter and is initialized with 0. By introducing α, the network starts from learning correlations around local regions, and then extends to learn the long-range dependency between different regions across the feature map.

Notably, in typical spatial attention, Queries and Keys are derived from the same feature map within a module. However, this paper's mechanism separates Queries from the Encoder module's feature map and Keys/Values from the Decoder module's feature map. This innovative method captures alignment and correlation between Encoder knowledge and Decoder predictions. A strong correlation results in a higher dot product, prompting the Decoder to extract more information from corresponding positions in its feature map. This introduces knowledge infusion from the Encoder, enhancing the Decoder's decision-making. Aligned Query and Key positions emphasize relevant Encoder knowledge, aiding accurate feature map construction. This unique spatial attention acts as a bridge, fostering collaboration between Encoder and Decoder through alignment-based learning.

D. LOSS FUNCTIONS

The proposed model is trained in a supervised manner. In the training process, we have used two types of loss functions: a) Fused Focal (FF) loss, and b) Dice loss. With the aim of having an end-to-end training process, we define the total loss L as:

$$L = L_{Fused} + L_{Dice} \tag{7}$$

1) FUSED FOCAL LOSS

Using FF loss, the class imbalance is mitigated (the areas that have been inpainted are often small when compared with the entire image). A FF loss function [38], [39] addresses class imbalance during training in tasks like object detection. The FF loss focuses on learning on hard misclassified examples by applying a modulating term to the cross-entropy loss. This is a dynamically scaled cross-entropy loss, in which the scaling factor becomes zero as confidence in the correct class increases. In a nutshell, this factor automatically down-weights the contribution of easy examples during training in order to focus the model on problems of difficulty more rapidly. However, in most of the inpainting-based forgeries, the inpainted areas are relatively smaller than the pristine ones, resulting in a class imbalance. As a consequence, the trained model tends to classify the samples as pristine more often. In order to address this issue, we propose to incorporate the FF loss into the binary cross-entropy loss, creating a FF loss function. A α-balanced variant of the FF loss is typically defined as follows:

$$L_{Fused} = -\sum_{i}^{n} \alpha(1-\hat{M}_i)^{\lambda} M_i \log \hat{M}_i + (1-\alpha)\hat{M}_i{}^{\lambda}(1-M_i)\log(1-\hat{M}_i) \tag{8}$$

where $\hat{M}_i$ and M_i are predicted output and grand-truth respectively, and n is the size of output vector. In particular, λ, is a focusing parameter that can smoothly adjust how easily examples are down-weighted. Clearly, when $\lambda = 0$, the focal loss is the same as the cross-entropy loss, and as λ increases, so does the impact of the modulating factor. We evaluate different choices of $\lambda \in (1, 2, 3)$, and empirically find that $\lambda = 2$ works best in the experiments. In addition, α is the weight assigned to the rare class for further adjusting imbalanced classes. We hence set $\alpha = 0.75$ to balance the rare class.

2) DICE LOSS

Dice loss [40] is widely used in medical image segmentation tasks to address the data imbalance problem. It only addresses the imbalance problem between foreground and background but overlooks another imbalance factors between easy and hard examples.

$$L_{Dice} = 1 - \frac{2\sum_{i}^{H\times W} \hat{M}_i \cdot M_i}{\sum_{i}^{H\times W} \hat{M}_i{}^2 + \sum_{i}^{H\times W} M_i^2} \tag{9}$$

IV. EXPERIMENTS AND DISCUSSION

In this section, we first introduce the experimental settings, then evaluate our proposed method on newly generated datasets based on Places2 [10] and CelebA [11]. We compare the results of our proposed method with two other recent methods called MVSS-Net [23] and LDICN [19]. For quantitatively measuring the performance difference among the methods, we utilize several statistical metrics. Finally, we report an ablation study on the effects of the residual block and the attention block in our proposed method.

A. TRAINING SETTING

We train the networks using the Adam optimizer with an initial learning rate of $1e^{-4}$. All of our experiments are run with a Nvidia Tesla P100 GPU.

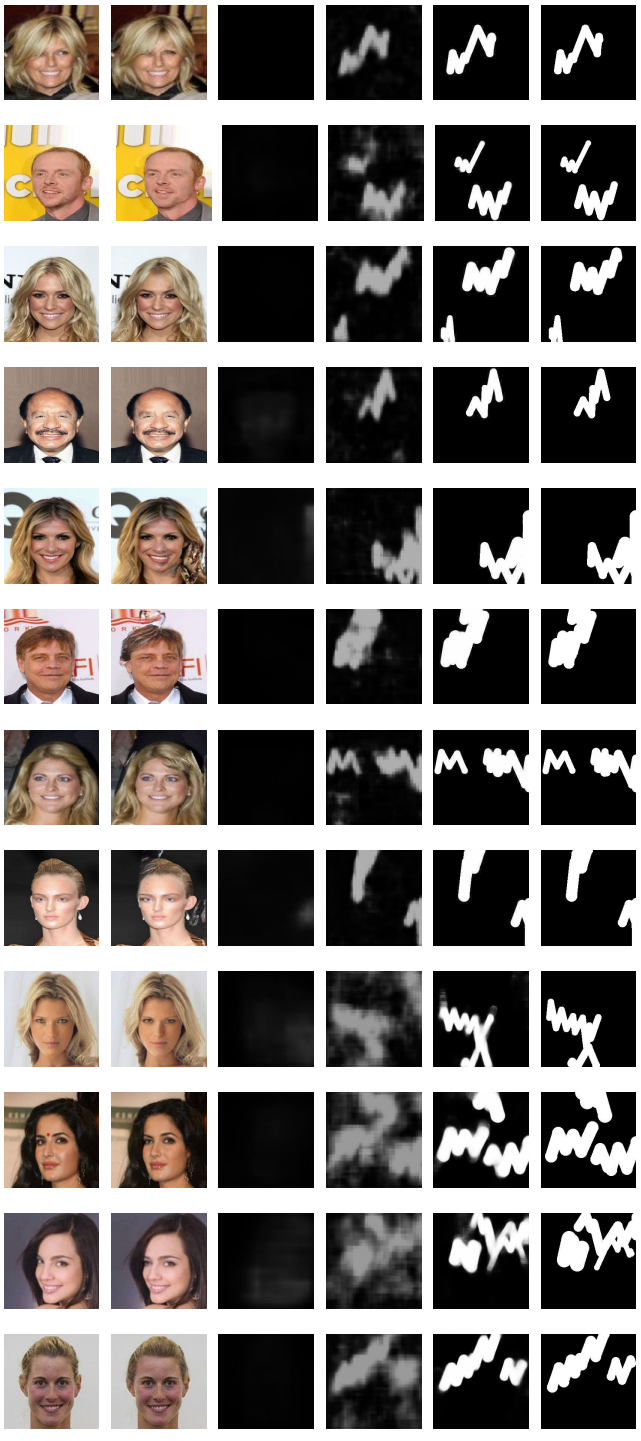


FIGURE 6. Quality comparison among our method and other methods on CelebA. The second column of each row shows the inpainted image using the mask in the last column (Ground Truth, GT), which is detected using the two reference methods (MVSS-Net and LDICN) and Ours.

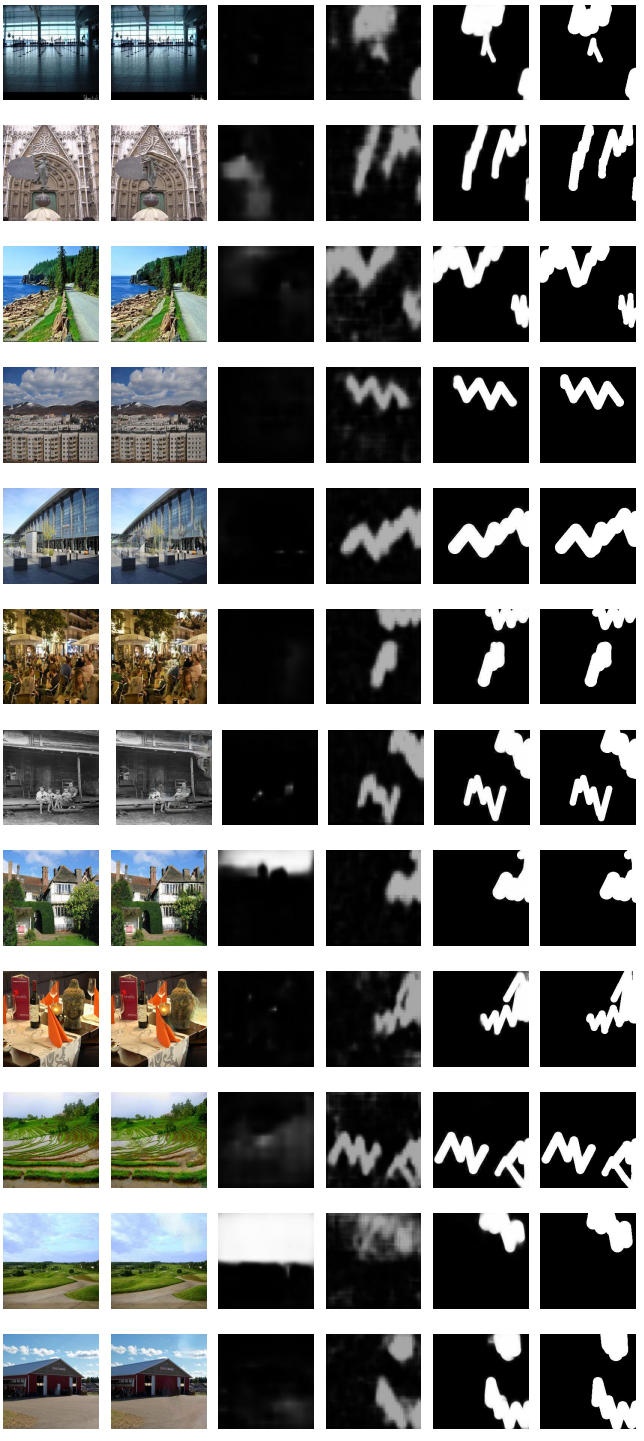


FIGURE 7. Quality comparison among our method and other methods on Places. The second column of each row shows theinpainted image using the mask in the last column (Ground Truth, GT), which is detected using the two reference methods (MVSS-Net and LDICN) and Ours.

B. DATASETS

We prepared the training and test data by exploiting Places [10] and CelebA [11] datasets. We used three different deep inpainting approaches, approaches including GC [12], CA [41], and EC [42] to generate inpainted images on the two mentioned datasets. For each of the two mentioned datasets, we randomly selected (without replacement) 50K and 10K images to create training and test subsets, respectively. For

TABLE 1. Quantitative results over Places2 dataset for IRL-Net and other compared methods (LDICN, MVSS-Net). The best result of each column is boldfaced.

Method	Inpainting Method	mIoU	F1 score	Precision	Recall
LDICN	GC	85.87	83.77	96.50	74.01
MVSS-Net		43.40	1.56	19.09	0.82
Ours		**90.62**	**92.86**	**98.89**	**87.76**
LDICN	CA	89.09	86.28	98.31	76.87
MVSS-Net		43.61	2.94	26.19	1.56
Ours		**91.76**	**93.72**	**99.59**	**88.51**
LDICN	EC	85.26	83.41	96.07	73.70
MVSS-Net		43.38	2.98	20.85	1.60
Ours		**90.20**	**92.39**	**98.20**	**87.23**

TABLE 2. Quantitative results over CelebA dataset for IRL-Net and other compared methods (LDICN, MVSS-Net). The best result of each column is boldfaced.

Method	Inpainting Method	mIoU	F1 score	Precision	Recall
LDICN	GC	85.87	83.77	96.50	74.01
MVSS-Net		43.42	1.65	21.37	00.85
Ours		**90.77**	**92.82**	**98.68**	**87.62**
LDICN	CA	88.63	86.01	98.02	76.62
MVSS-Net		43.40	2.21	37.06	1.14
Ours		**91.87**	**93.81**	**99.68**	**88.58**
LDICN	EC	76.76	73.60	89.96	62.28
MVSS-Net		43.33	2.08	19.14	1.10
Ours		**87.50**	**89.44**	**95.91**	**83.79**

inpainting images, mask generation is an important factor. To simulate more diverse and complex real-world scenarios, we utilize the irregular mask setting in GC with arbitrary shapes and random locations for both training and testing. We underscore that this approach enables us to attain masks with diverse shapes and positions, thereby promoting variability in both our training and testing stages. Hence, our generated datasets contain tuples of the inpainted image and generated mask. (See Figs. 6 for example masks.).

C. EVALUATION METRICS

Four commonly used pixel-wise classification metrics, including Recall, Precision, mean Intersection over Union (mIoU), and F1-score, are adopted to evaluate the performance. The metrics are calculated on each image independently, and the mean values obtained over all images are reported in the following experiments. The mIoU metric is preferred since it is not affected by imbalanced classes. The Precision metric demonstrates how many instances that have been predicted true are really true. Meanwhile, Recall shows how many true positive instances are predicted correctly. F1-score is also used to combine the Precision and Recall metrics into a single metric.

D. COMPARISON WITH PREVIOUS WORK

In this section, we compare our proposed method with two state-of-the-art methods. However, for a fair comparison, we consider two main criteria to choose an appropriate state-of-the-art: a) Pre-trained models released by paper authors, and b) Source code publicly available. Accordingly, we have chosen MVSS-Net [23] and LDICN [19] for a fair comparison. MVSS-Net [23] was pre-trained on the CASIAv2 and DEFACTO datasets. We trained LDICN [19] again using the code provided by the authors on our datasets.

Therefore, we have two detection networks (LDICN [19] and MVSS-Net [23]), six training datasets created using three inpainting methods (GC [12], CA [41], and EC [42]), and two testing datasets (Places2 [10] and CelebA [11]). All of our experiments are run separately for Places2 and CelebA. The performance of the two reference detection networks and our proposed detection network are shown in Table 1 and Table 2.

1) QUANTITATIVE PERFORMANCE EVALUATION

As shown in Table 1 and Table 2, our proposed method outperforms existing methods by a large margin in all test scenarios. In the following, we will provide a detailed analysis of these results.

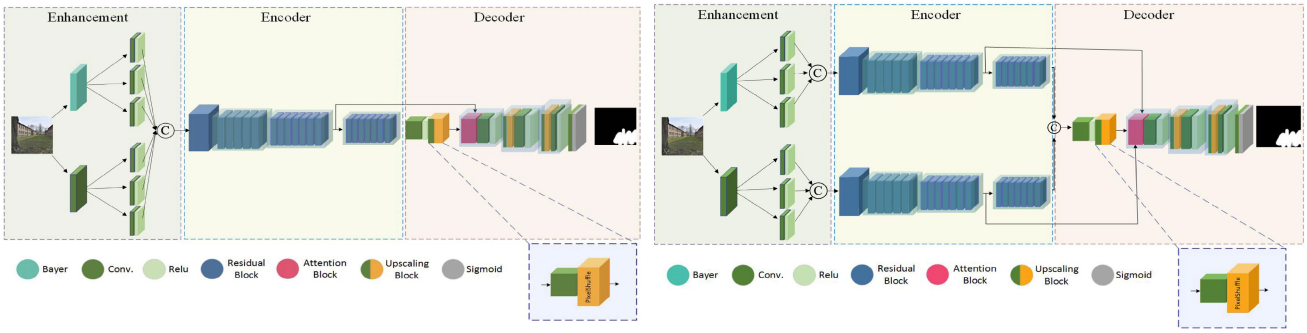
The detection results are reasonably good for LDICN (retrained on our datasets), but anyway worse than the ones obtained by our proposed ILR-Net. On the other hand, MVSS-Net is pre-trained on CASIAv2 and is reported to have a very good performance on that dataset, but its performance drops drastically here on our inpainting methods and datasets. Such poor generalizability indicates that MVSS-Net tends to overfit focusing on the artifacts of a particular inpainting method and fails to consider the common characteristics of different inpainting techniques. This is a common problem in image manipulation detection, the lack of generalization capabilities [16]. As discussed before, this generalization can be improved by properly exploiting the noise information contained in real versus inpainted contents. This also indicates that noise patterns are indeed a reliable cue for detecting inpainted regions.

2) QUALITATIVE PERFORMANCE EVALUATION

Using visuals, we present a qualitative comparison of the detected masks. Figs. 6 and 7 illustrate examples of LDICN and MVSS-Net using Places2 and CelebA. LDICN and MVSS-Net, however, cannot accurately identify the inpainted regions, especially when they are complex. Our proposed method achieves very good results on test samples of different inpainting methods.

E. FUSION EXPERIMENT

Our proposed method uses two kinds of information: noise-based information and RGB-based information, corresponding respectively to the Bayar and Convolution layers in Fig. 2 right after the input image. It is important to combine and



(a) Middle fusion architecture

(b) Late fusion architecture

FIGURE 8. The proposed middle and late fusion architectures.

TABLE 3. Comparison of three fusion techniques.

Architecture	mIoU	F1-score	Recall	Precision
early-fusion	**90.62**	**92.86**	87.76	**98.89**
middle-fusion	89.95	92.71	**88.08**	97.85
late-fusion	90.14	92.55	87.59	98.12

TABLE 4. Impact of SE-Residual and attention blocks on performance.

Architecture	mIoU	F1-score	Recall	Precision
w BN	89.49	92.62	**88.29**	97.39
w/o BN	89.80	92.23	87.20	97.88
w/o BN w SE	90.12	92.15	86.71	98.31
w Attention	**90.62**	**92.86**	87.76	**98.89**

fuse this information at a specific stage in order to carry out further processing. Note that, in Fig. 2, that fusion is implemented in the Enhancement module via a concatenation operation in a kind of early fusion. Here, we analyze other architectures for combining the RGB and noise information. For this purpose, we consider three types of fusion: early fusion, middle fusion, and late fusion. We used early fusion in our proposed method represented in Fig. 2 combining the mentioned information right after the first layer. For middle fusion (see Fig. 8 (a)), more information is extracted from the noise-based and RGB-based channels. Thus, in the Enhancement stage, we have two branches that extract information separately, and at the end of the Enhancement stage, we combine the extracted information. For late-fusion (see Fig. 8 (b)), we use two independent branches based on RGB and noise information all throughout the Enhancement and Encoder stages. In this case, we place the fusion function (again feature concatenation) after the Encoder stage. Table 3 shows the measured quantitative results for the three fusion methods. As can be observed, the early-fusion technique outperforms the other two fusion methods.

F. ABLATION STUDY: RESIDUAL AND ATTENTION UNITS

In this section, we quantitatively analyze the impact of the proposed SE-Residual and Attention blocks in our model. As shown in Table 4, each of the two components in IRL-Net contributes in different ways to its performance. By comparing the statistical results presented in Table 4 with those of previous residual blocks, our suggested residual block extracts the necessary information quite well. The attention block combines asymmetrically two separate feature embeddings of the same dimension, in contrast, the self-attention input is a single feature embedding. The quantitative results demonstrate the advantages of the attention block.

V. CONCLUSION

To improve the performance of detecting and localizing manipulated image regions, this paper has proposed a novel method, named IRL-Net (Inpainted Region Localization Network). The proposed method uses high-level features extracted from both the RGB image and a high-pass filtered version of the RGB image concatenated at some stage for further processing. It also performs end-to-end training to learn the discriminative features between manipulated and non-manipulated regions through back-propagation using ground truth and masked image. IRL-Net consists of two important feature designs: a) a new Residual block based on Squeeze-and-Excitation, and b) an Attention block combines the two feature embeddings according to their information. IRL-Net achieves promising results in localizing manipulated regions at pixel level on testing datasets. Future work includes analyzing and improving the generalization capabilities specially against unseen manipulations [16], [43], and further exploration of more sophisticate fusion architectures [27] combining image contents and noise elements.

ACKNOWLEDGMENT

(Amir Etefaghi Daryani, Mahdieh Mirmahdi, and Ahmad Hassanpour contributed equally to this work.)

REFERENCES

[1] Y. Rao and J. Ni, "A deep learning approach to detection of splicing and copy-move forgeries in images," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2016, pp. 1–6, doi: 10.1109/WIFS.2016.7823911.

[2] Y. Wu, W. Abd-Almageed, and P. Natarajan, "BusterNet: Detecting copy-move image forgery with source/target localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1–17.

[3] M. Aria, M. Hashemzadeh, and N. Farajzadeh, "QDL-CMFD: A quality-independent and deep learning-based copy-move image forgery detection method," *Neurocomputing*, vol. 511, pp. 213–236, Oct. 2022.
[4] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector," in *Proc. IEEE Int. WIFS*, Nov. 2015, pp. 1–6, doi: 10.1109/WIFS.2015.7368565.
[5] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, vol. 11215.
[6] V. V. Kniaz, V. Knyaz, and F. Remondino, "The point where reality meets fantasy: Mixed adversarial generators for image splice detection," in *Proc. NeurIPS*, vol. 32, 2019, pp. 1–12.
[7] M.-J. Kwon, I.-J. Yu, S.-H. Nam, and H.-K. Lee, "CAT-net: Compression artifact tracing network for detection and localization of image splicing," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 375–384.
[8] H. Wu and J. Zhou, "IID-net: Image inpainting detection network via neural architecture search and attention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1172–1185, Mar. 2022.
[9] H. Xiang, Q. Zou, M. A. Nawaz, X. Huang, F. Zhang, and H. Yu, "Deep learning for image inpainting: A survey," *Pattern Recognit.*, vol. 134, Feb. 2023, Art. no. 109046.
[10] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[11] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. ICCV*, 2015, pp. 3730–3738.
[12] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4470–4479.
[13] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "An introduction to digital face manipulation," in *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Cham, Switzerland: Springer, 2022, pp. 3–26.
[14] R. Tolosana, C. Rathgeb, R. Vera-Rodriguez, C. Busch, L. Verdoliva, S. Lyu, H. H. Nguyen, J. Ya-Magishi, I. Echizen, and P. Rot, "Future trends in digital face manipulation and detection," in *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Cham, Switzerland: Springer, 2022, pp. 463–482.
[15] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and Y. Akbari, "Image inpainting: A review," *Neural Process. Lett.*, vol. 51, no. 2, pp. 2007–2028, 2019.
[16] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 1038–1048, Aug. 2020.
[17] 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.
[18] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2691–2706, Nov. 2018.
[19] H. Li and J. Huang, "Localization of deep inpainting using high-pass fully convolutional network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8301–8310.
[20] C. Yang, H. Li, F. Lin, B. Jiang, and H. Zhao, "Constrained R-CNN: A general image manipulation detection model," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
[21] X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, and R. Nevatia, "SPAN: Spatial pyramid attention network for image manipulation localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 312–328.
[22] Y. Wu, W. AbdAlmageed, and P. Natarajan, "ManTra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proc. CVPR*, 2019, pp. 1–10.
[23] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, "MVSS-net: Multi-view multi-scale supervised networks for image manipulation detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3539–3553, Mar. 2023.
[24] R. Salloum, Y. Ren, and C.-C. Jay Kuo, "Image splicing localization using a multi-task fully convolutional network (MFCN)," *J. Vis. Commun. Image Represent.*, vol. 51, pp. 201–209, Feb. 2018.
[25] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1053–1061.
[26] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. S. Manjunath, "Exploiting spatial structure for localizing manipulated image regions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4980–4989.
[27] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics. Part 1: Fundamentals and review," *Inf. Fusion*, vol. 44, pp. 57–64, Nov. 2018.
[28] A. Hassanpour, A. E. Daryani, M. Mirmahdi, K. Raja, B. Yang, C. Busch, and J. Fierrez, "E2F-GAN: Eyes-to-face inpainting via edge-aware coarse-to-fine GANs," *IEEE Access*, vol. 10, pp. 32406–32417, 2022.
[29] X. Zhang, D. Zhai, T. Li, Y. Zhou, and Y. Lin, "Image inpainting based on deep learning: A review," *Inf. Fusion*, vol. 90, pp. 74–94, Feb. 2022.
[30] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proc. ICCV*, 2019, pp. 4170–4179.
[31] Z. Guo, Z. Chen, T. Yu, J. Chen, and S. Liu, "Progressive image inpainting with full-resolution residual network," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2496–2504.
[32] P. Zhou, B.-C. Chen, X. Han, M. Najibi, A. Shrivastava, S.-N. Lim, and L. Davis, "Generate, segment, and refine: Towards generic manipulation segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 13058–13065, Apr. 2020.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[34] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[35] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, 2017, pp. 4681–4690.
[36] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 257–265.
[37] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[39] J. Chang, X. Zhang, J. Chang, M. Ye, D. Huang, P. Wang, and C. Yao, "Brain tumor segmentation based on 3D UNet with multi-class focal loss," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2018, pp. 1–5.
[40] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. DLMIA ML-CDS*, 2017, pp. 240–248.
[41] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.
[42] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative image inpainting with adversarial edge learning," 2019, *arXiv:1901.00212*.
[43] P. Korshunov and S. Marcel, "Improving generalization of deepfake detection with data farming and few-shot learning," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 3, pp. 386–397, Jul. 2022.
[44] R. Ren, Q. Hao, S. Niu, K. Xiong, J. Zhang, and M. Wang, "MFI-net: Multi-feature fusion identification networks for artificial intelligence manipulation," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jun. 26, 2023, doi: 10.1109/TCSVT.2023.3289171.
[45] H. Ding, L. Chen, Q. Tao, Z. Fu, L. Dong, and X. Cui, "DCU-net: A dual-channel U-shaped network for image splicing forgery detection," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 5015–5031, Mar. 2023.
[46] R. Ren, S. Niu, J. Jin, J. Zhang, H. Ren, and X. Zhao, "Multi-scale attention context-aware network for detection and localization of image splicing: Efficient and robust identification network," *Appl. Intell.*, vol. 53, no. 15, pp. 18219–18238, 2023.

[47] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, "F2Trans: High-frequency fine-grained transformer for face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1039–1051, 2023.
[48] A. K. Jaiswal, S. Singh, S. K. Tripathy, N. K. Tagore, and A. Shahi, "OME methods for digital image forgery detection and localization," in *Proc. Data Analytics Manag. (ICDAM)*, Singapore, 2022, pp. 226–245.
[49] Q. Zeng, H. Wang, Y. Zhou, R. Zhang, and S. Meng, "A parallel attention mechanism for image manipulation detection and localization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[50] Y. Xu, M. Irfan, A. Fang, and J. Zheng, "Multiscale attention network for detection and localization of image splicing forgery," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023.

AMIR ETEFAGHI DARYANI (Student Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Guilan, in 2019. He is currently pursuing the M.Sc. degree in electrical engineering with the Amirkabir University of Technology. He was a Research Assistant with the Digital Systems Laboratory, Amirkabir University of Technology, which was directed by Prof.Saeed Sharifian to pursue the M.Sc. degree. His research interests include the development of decision-making systems based on the reinforcement learning approaches and image processing by using deep learning-based techniques to build autonomous systems that act like humans.

MAHDIEH MIRMAHDI (Member, IEEE) received the B.Sc. degree in computer science and the M.Sc. degree in computer engineering-artificial intelligence and robotics from the University of Isfahan, in 2017 and 2021, respectively. During the master's degree, she conducted research in the field of computer vision, specifically focusing on video semantic segmentation under the guidance of Prof. Amirhassan Monadjemi. Her research interests include machine/deep learning and computer vision.

AHMAD HASSANPOUR (Student Member, IEEE) received the M.Sc. degree in computer engineering from the Shiraz University of Technology. He is currently pursuing the Ph.D. degree with the Department of Information Security and Communication Technology (IIK), Norwegian University of Science and Technology (NTNU), Norway. He is also a Marie Skłodowska-Curie Fellow (H2020 Privacy Matters Project). His research interests include deep learning, computer vision, and privacy.

HATEF OTROSHI SHAHREZA (Graduate Student Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering from the University of Kashan, Iran, in 2016, and the M.Sc. degree in electrical engineering from the Sharif University of Technology, Iran, in 2018. He is currently pursuing the Ph.D. degree with École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. He is also a Research Assistant with the Biometrics Security and Privacy Group, Idiap Research Institute, Switzerland, where he received the H2020 Marie Skłodowska-Curie Fellowship (TReSPAsS-ETN) for his doctoral program. During the Ph.D. degree, he also experienced being a Visiting Scholar with the Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, for six months. He is also the Winner of the European Association for Biometrics (EAB) Research Award 2023. His research interests include deep learning, machine learning, computer vision, biometrics, and biometric template protection.

BIAN YANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic engineering and information security from the Harbin Institute of Technology, in 2000, 2002, and 2006, respectively. He visited Fraunhofer IGD, Darmstadt, from 2003 to 2005, for research on data hiding and media security, as a Lecturer with the School of Computer Science and Technology, Harbin Institute of Technology, from 2005 to 2007, and a Research Engineer with Thomson Corporate Research, Beijing, from 2007 to 2008. He joined the Norwegian Information Security Laboratory (NISlab), Gjøvik University College, working on privacy-preserving biometrics, from 2008 to 2015, and founded, in 2016. He has been coordinating the eHealth and Welfare Security (eHWS) Group, Department of Information Security and Communication Technology, Norwegian University of Science and Technology (NTNU). His research interests include cybersecurity and privacy for e-health and welfare technologies and services, privacy modeling and enhancing technologies, security biometrics and identity management, and security practice and human factors.

JULIAN FIERREZ (Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications engineering from Universidad Politécnica de Madrid, Spain, in 2001 and 2006, respectively. Since 2004, he has been with Universidad Autónoma de Madrid, where he is currently an Associate Professor. From 2007 to 2009, he was a Visiting Researcher with Michigan State University, USA, under a Marie Curie Postdoctoral. His research interests include signal and image processing, HCI, responsible AI, and biometrics for security and human behavior analysis. He is actively involved in large EU projects in these topics (e.g., TABULA RASA and BEAT in the past and currently IDEA-FAST and TRESPASS-ETN), and has attracted notable impact for his research. He is a member of the ELLIS Society. He was a recipient of a number of distinctions, including the EAB Industry Award, in 2006, the EURASIP Best Ph.D. Award, in 2012, and the 2017 IAPR Young Biometrics Investigator Award. He has received best paper awards at ICB and ICPR. He is an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and the IEEE TRANSACTIONS ON IMAGE PROCESSING.