Discrepancy-Guided Reconstruction Learning for Image Forgery Detection

Zenan Shi[1,2] , **Haipeng Chen**[1,2] , **Long Chen**[3] , **Dong Zhang**[3,*]
[1]College of Computer Science and Technology, Jilin University
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University
[3]Department of CSE, The Hong Kong University of Science and Technology
{shizn, chenhp}@jlu.edu.cn, {longchen, dongz}@ust.hk

Abstract

In this paper, we propose a novel image forgery detection paradigm for boosting the model learning capacity on both forgery-sensitive and genuine compact visual patterns. Compared to the existing methods that only focus on the discrepant-specific patterns (*e.g.*, noises, textures, and frequencies), our method has a greater generalization. Specifically, we first propose a Discrepancy-Guided Encoder (DisGE) to extract forgery-sensitive visual patterns. DisGE consists of two branches, where the mainstream backbone branch is used to extract general semantic features, and the accessorial discrepant external attention branch is used to extract explicit forgery cues. Besides, a Double-Head Reconstruction (DouHR) module is proposed to enhance genuine compact visual patterns in different granular spaces. Under DouHR, we further introduce a Discrepancy-Aggregation Detector (DisAD) to aggregate these genuine compact visual patterns, such that the forgery detection capability on unknown patterns can be improved. Extensive experimental results on four challenging datasets validate the effectiveness of our proposed method against state-of-the-art competitors.

1 Introduction

Advanced image co-editing and synthesis methods make it cushy for people to tamper with images [Zhu *et al.*, 2020; Rombach *et al.*, 2022]. For example, objects and external properties of these objects in a given image can be completely interpolated via a few texts [Kawar *et al.*, 2022]. Although these progressive methods can increase the diversity and interest of images, on the other hand, they cause a new problem that people's confidence in the information expressed in images is reduced [Cao *et al.*, 2022; Fei *et al.*, 2022]. Besides, tampered images may also be used in some malicious occasions (*e.g.*, fake news and deliberate slanders), thus bringing potential social harms [Hu *et al.*, 2021; Zhang *et al.*, 2022a; Sun *et al.*, 2022b; Li *et al.*, 2020a]. Therefore, exploring effective image forgery detection methods is urgent.

In recent years, thanks to the immense progress of image processing technologies based on the deep learning mechanism [He *et al.*, 2016; Zhang *et al.*, 2020], ample semantic features greatly improve the recognition accuracy of image forgery detection [Wang and Deng, 2021; Zhuang *et al.*, 2022] on both the image-level and the pixel-level [Jiang *et al.*, 2020; Sun *et al.*, 2021; Zhao *et al.*, 2021]. However, off-the-shelf deep learning methods cannot achieve satisfactory results in the face of some challenging forgery cases (*e.g.*, unusual tampering areas and marginal tampering clues). To address this problem and improve the accuracy, some recent approaches use specific operators (*e.g.*, BayerConv [Chen *et al.*, 2021b], Sobel operator [Chen *et al.*, 2021b], and frequency filter [Fei *et al.*, 2022]) to extract discrepant-specific patterns (*e.g.*, noises, textures, and frequencies) as a supplementary for the recognition model. What these methods have commonly is that they use a mainstream backbone network and an additional auxiliary network to extract implicit and explicit semantic features, respectively [Sun *et al.*, 2022b; Wang and Deng, 2021; Zhuang *et al.*, 2022; Chen *et al.*, 2021b; Zhao *et al.*, 2021]. However, these methods usually need to generate temporary supervisions via intermediate feature maps, which are inherently not accurate enough, thus hurting the recognition effectiveness.

What's more, the generalization capacity of existing methods is somewhat limited – due to their overemphasis on the consequence of the explicit discrepant (*i.e.*, the tampered region) features, which limits their latent usage scope. To be specific, only learning some specific types of tampering patterns is far from pragmatism, because we cannot suppose tampering manners [Cao *et al.*, 2022; Yoshihashi *et al.*, 2019; Chen *et al.*, 2021a]. To improve the generalization capacity, it is helpful to learn a set of compact visual patterns, which inherently contain some general image properties, *e.g.*, the concurrent local textures, the consistent regional resolutions, and the continuous bright changes [Robert *et al.*, 2018; Cao *et al.*, 2022; Yoshihashi *et al.*, 2019]. For image forgery detection, to achieve this goal, some work demonstrated that image reconstruction is an effective approach [Wang *et al.*, 2022b; Li *et al.*, 2022b]. The reconstructed output has rich compact patterns and suppresses local forgery regions. However, the existing methods are usually equipped with a single reconstruction head, which suffers from problems of tedious feature representations and inadequate reasoning ability.

*Corresponding author: Dong Zhang.

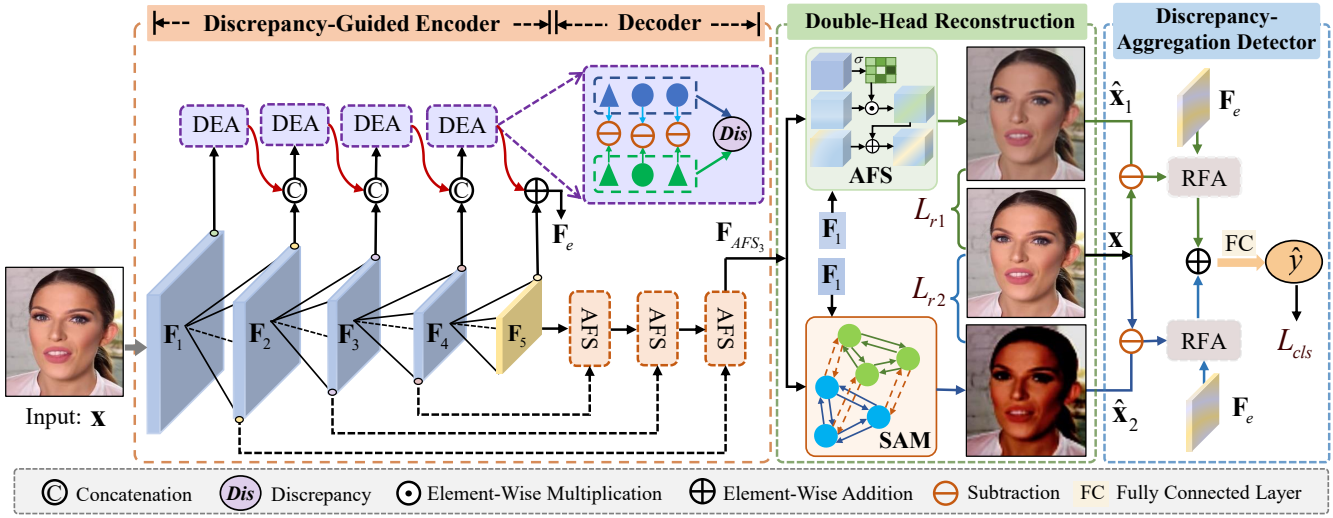

Figure 1: Overall architecture of our DisGRL, which mainly consists of: a Discrepancy-Guided Encoder (DisGE), a decoder, a Double-Head Reconstruction (DouHR) module, and a Discrepancy-Aggregation Detector (DisAD) head network for image forgery classification.

In this paper, we propose a novel image forgery detection paradigm, named Discrepancy-Guided Reconstruction Learning (DisGRL), to improve the model learning capacity on both the forgery-sensitive and the genuine compact visual patterns. As illustrated in Figure 1, DisGRL consists of four components: a Discrepancy-Guided Encoder (DisGE), a decoder, a Double-Head Reconstruction (DouHR) module, and a Discrepancy-Aggregation Detector (DisAD) head network.

Specifically, the proposed DisGE (*ref.* Sec. 3.1) is used to extract forgery-sensitive visual patterns, which consists of two branches: a mainstream backbone branch is used to extract the general semantic features, and an accessorial discrepant external attention branch is used to extract the explicit forgery visual cues. Thereby, DisGE is more suitable for the image forgery detection task than the common backbone networks (*e.g.*, convolutional neural networks and vision transformer). In the decoder, three progressive attention feature selection modules are employed in $\mathbf{F}_i$ to connect feature maps from the corresponding encoder network layer, which finally has the same scale as $\mathbf{F}_2$. We also propose a DouHR module (*ref.* Sec. 3.2) based on the decoder, which can enhance the genuine compact visual patterns in two separate granular spaces via an image reconstruction manner. To be specific, in the DouHR module, an Attention-guidance Feature Selection (AFS) procedure and a Similarity Aggregation Module (SAM) are used to extract the vision-based and the reasoning-based genuine compact visual patterns, respectively. Based on the DouHR, we further introduce a DisAD head network (*ref.* Sec. 3.3) for image forgery classification, which can aggregate the obtained genuine compact visual patterns via a Reconstruction-guidance Feature Aggregation (RFA) module, resulting in an improved forgery detection capability on unknown patterns. Therefore, compared to the existing methods that only focus on these discrepant-specific patterns, our proposed DisGRL has a stronger generalization capability. To demonstrate the superiority of DisGRL, extensive experiments are carried out on four commonly used yet challenging face forgery detection datasets. Results validate that our DisGRL can achieve state-of-the-art performance on both seen and unseen forgeries.

Our contributions are as follows: 1) We propose a novel DisGRL for image forgery detection, which contains three proposed components for learning both forgery-sensitive and genuine compact visual patterns. 2) Extensive experimental results on four challenging datasets validate that DisGRL can achieve state-of-the-art performance against competitors.

2 Related Work

Image forgery can be viewed as a game of AI *v.s.* AI since the majority of detection technologies are based on deep learning. In the past, many efforts have been made to improve the performance of natural/face image forgery detection [Fei *et al.*, 2022; Haliassos *et al.*, 2021; Gu *et al.*, 2022; Zhang *et al.*, 2021]. Extensive work uses a two-branch architecture to mine specific forgery patterns, such as noises or frequency domain features in combination with RGB spatial data, in light of the fact that altered images are getting more visually realistic [Li *et al.*, 2022a; Chen *et al.*, 2021a; Masi *et al.*, 2020; Qian *et al.*, 2020; Li *et al.*, 2021; Wang *et al.*, 2022a]. SOLA [Jia *et al.*, 2022] fuses multimodal features from RGB and high-frequency features extracted by a DCT transformation in an extra branch for more general representations. As a complementary of RGB, the model in [Fei *et al.*, 2022] introduces subtle noise features via learnable high pass filters with anomalies in local regions also performed well in unseen forgeries [Zhang *et al.*, 2020; Yan *et al.*, 2023; Zhang *et al.*, 2022a]. Despite their remarkable performance, their models for obtaining specific forgery patterns only reflect certain aspects of the forgery, which might lead to model bias or sub-optimization.

Recently, some advanced methods are proposed to improve the model generalization capacity such as exploiting con-

trastive learning to guide the recognition model focus on local content inconsistencies [Sun *et al.*, 2022b; Shi *et al.*, 2023; Zhang *et al.*, 2022b], introducing domain adaptation to alleviate overfitting on a single domain [Rao *et al.*, 2022; Sun *et al.*, 2021; Rao and Ni, 2021], and/or enhance feature representation with an information-theoretic self-information metric for forgery detection [Sun *et al.*, 2022a]. These methods achieve both great performances under intra-dataset (*i.e.*, seen) and cross-domain (*i.e.*, unseen) evaluations. Unlike these methods that explore the local level inconsistencies, our method focuses more on forgery-sensitive and genuine compact visual patterns, which can improve both model's accuracy and generalization.

3 Our Approach

DisGRL is proposed to improve the model learning capacity on both the forgery-sensitive and the genuine compact visual patterns. Our contributions lie in presenting: a Discrepancy-Guided Encoder (DisGE), a Double-Head Reconstruction (DouHR) module, and a Discrepancy-Aggregation Detector (DisAD) head network for image forgery classification. An overview architecture of DisGRL is illustrated in Figure 1. The input is an RGB image $\mathbf{X}$, and the output is a binary predicted label $\hat{y}$, which indicates whether the input image is forged or not. In the following, we detail the implementations of each proposed component.

3.1 Discrepancy-Guided Encoder (DisGE)

To capture forgery-sensitive visual patterns, we propose a DisGE, which consists of two parallel branches, where the mainstream backbone based on Xception network [Chollet, 2017] is used to extract multi-level semantic features, *i.e.*, $\mathbf{F}_i$ $(i = 1, 2, \ldots, 5)$, and the **Discrepancy External Attention (DEA)** branch is applied to different level feature to extract explicit discrepant-specific pattern, which are usually subtle and occur in local regions. As shown in Figure 1, features from different Xception layers are combined in a cascaded manner by DEA block. The specific operation of each DEA block's output $\mathbf{D}_i$ is expressed as:

$$\mathbf{D}_i = \begin{cases} Dea(\mathbf{F}_i), & i = 1 \\ Dea(Cat(\mathbf{D}_{i-1}, \mathbf{F}_i)), & i \in [2, 3, 4] \end{cases} \quad (1)$$

where $Dea(\cdot)$ and $Cat(\cdot)$ denote each DEA block and feature concatenation along the channel dimension, respectively.

For each DEA block, as shown in Figure 2, we first apply a 3×3 convolutional layer on the input feature maps $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ with the same channel size C. Then, an adaptive average pooling is used to obtain the pooled features $\mathbf{F}_d$. After that, the differentiated maps can be obtained through $\mathbf{D}' = \mathbf{F} - \mathbf{F}_d$ to extract the discrepant information. Inspired by [Guo *et al.*, 2023], two 1D convolutions that share the same parameters are further introduced to characterize the global features of the entire map. Concretely, given a differentiated input $\mathbf{D}' \in \mathbb{R}^{C \times H \times W}$, after reshaping and 1D convolution, feature maps are up-sampled four times in channel size. And 1D convolution and a reshape function are applied again to restore the original feature map size. Finally, the output feature map $\mathbf{D}$ can be obtained through a 1×1 convolution and a residual connection.

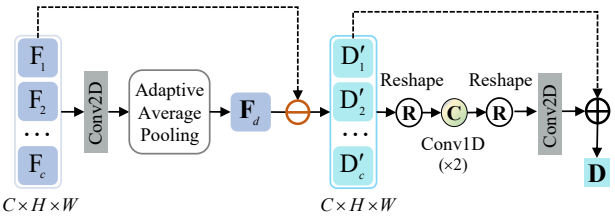

Figure 2: Illustration of the proposed Discrepancy External Attention (DEA) block, which is proposed to extract the forgery-sensitive visual patterns in the Discrepancy-Guided Encoder network.

3.2 Double-Head Reconstruction (DouHR)

Reconstruction learning has been proven to be beneficial to several image forgery detection works by exploring rich compact visual patterns [Cao *et al.*, 2022; Wang *et al.*, 2022b; Li *et al.*, 2022b]. In this work, we propose DouHR module based on the decoder to enhance the genuine compact visual patterns in two separate granular spaces (*i.e.*, AFS and SAM) via an image reconstruction manner, such that the model can not only learn a rich genuine compact visual pattern but also further suppress the visual representation of the local forgery regions. As shown in Figure 1, besides $\hat{\mathbf{X}}_1$ that is generated by an **Attention-guidance Feature Selection (AFS)** procedure in extracting the vision-based genuine compact visual patterns via convolutions, we introduce an extra **Similarity Aggregation Module (SAM)** for extracting the reasoning-based genuine compact visual patterns via secondary reconstruction $\hat{\mathbf{X}}_2$. The DouHR module can be formulated as:

$$\hat{\mathbf{X}}_1 = \mathrm{AFS}(\mathbf{F}_{AFS_3}, \mathbf{F}_1), \hat{\mathbf{X}}_2 = \mathrm{SAM}(\mathbf{F}_{AFS_3}, \mathbf{F}_1), \quad (2)$$

where $\mathbf{F}_{AFS_3}$ indicates the output feature maps of the third AFS procedure in the decoder. In DouHR, we adjust the number of channels from the output of the SAM and AFS modules to 3 by applying a 1×1 convolution. After that, we use bilinear interpolation to adjust the feature map size to match the input image size.

AFS. In the decoder, three AFS modules receive the output of the previous AFS module and feature maps of the corresponding level in the mainstream backbone as input. For example, the inputs to the third AFS are $\mathbf{F}_{AFS_2}$ and $\mathbf{F}_2$. In DouHR, AFS receives the $\mathbf{F}_{AFS_3}$ and as $\mathbf{F}_1$ input. The concatenating operation $Cat(\cdot)$ is first carried out on $\mathbf{F}_{AFS_3}$ and $\mathbf{F}_1$ in the channel dimension, *i.e.*, $\tilde{\mathbf{F}} = Cat(\mathbf{F}_{AFS_3}, \mathbf{F}_1)$, followed by a depthwise separable convolution f_{d3} to obtain attention map $\mathbf{A}_{att}$ with the same shape as input features and suppress the unimportant region of feature information transmitted by decoder output, so that model pays more attention to the genuine compact visual patterns. Finally, a residual connection operation is applied to obtain the output. The above process can be expressed as follows:

$$\begin{aligned} \mathbf{A}_{att} &= \sigma(f_{d3}(\tilde{\mathbf{F}})), \\ \mathbf{A} &= f_{d3}(f_{d3}(\tilde{\mathbf{F}}) \odot \mathbf{A}_{att}) + f_{c3}(\tilde{\mathbf{F}}), \end{aligned} \quad (3)$$

where f_{c3} and $\sigma(\cdot)$ are the 3×3 convolution layer and sigmoid activaion function, respectively. Other AFS procedures are calculated in a similar way.

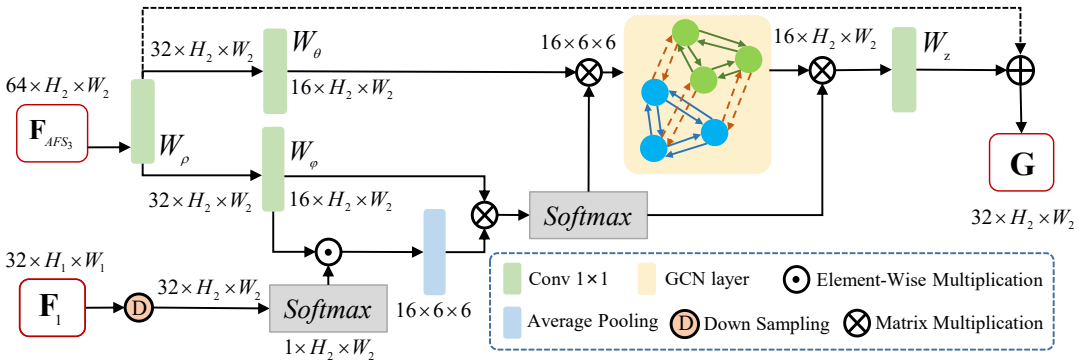

Figure 3: Illustration of Similarity Aggregation Module (SAM), which can extract the reasoning-based genuine compact visual patterns.

SAM. To inject detailed global features into high-level semantic features using a global reasoning reconstruction, inspired by [Dong *et al.*, 2021], we introduce non-local operation under graph convolution operation [Lu *et al.*, 2019; Zhang *et al.*, 2022b] to implement SAM. As shown in Figure 3, for the given feature map $\mathbf{F}_{AFS_3}$, we first apply three 1×1 convolutions (*i.e.*, $\mathbf{W}_\rho$, $\mathbf{W}_\theta$, and $\mathbf{W}_\varphi$) to reduce the channel dimension into 16, and obtain feature maps $\mathbf{F}_\theta$, $\mathbf{F}_\varphi$, which can be expressed as:

$$\mathbf{F}_\theta = \mathbf{W}_\theta(\mathbf{W}_\rho(\mathbf{F}_{AFS_3})),\ \mathbf{F}_\varphi = \mathbf{W}_\varphi(\mathbf{W}_\rho(\mathbf{F}_{AFS_3})). \quad (4)$$

For $\mathbf{F}_1$, we down-sample it to the same size as $\mathbf{W}_\rho(\mathbf{F}_{AFS_3})$. Then we apply a Softmax function along the channel dimension and calculate the element-wise multiplication with $\mathbf{F}_\varphi$ for assigning different weights to different pixels and increasing the weight of edge pixels. And an adaptive pooling operation $Avp(\cdot)$ is utilized to reduce the displacement of features. In summary, the processing can be formulated as:

$$\mathbf{F_w} = Avp(\mathbf{F}_\varphi \odot \mathrm{softmax}(D(\mathbf{F}_1))), \quad (5)$$

where $D(\cdot)$ and $\mathrm{softmax}(\cdot)$ denote the down-sampling and Softmax functions, respectively. After that, the matrix multiplication and Softmax function are used to establish the correlation between $\mathbf{F}_\varphi$ and $\mathbf{F_w}$, which can be expressed as:

$$\mathbf{F_{cor}} = \mathrm{softmax}(\mathbf{F_w} \otimes (\mathbf{F}_\varphi)^T). \quad (6)$$

The correlation attention map $\mathbf{F_{cor}}$ is multiplied with the feature map $\mathbf{F}_\theta$, and the resulting map is fed to the graph convolutional newtwork (GCN). Same to [Dong *et al.*, 2021], reconstructing the graph domain features into the original structural features as follows:

$$\mathbf{G}' = \mathbf{F_{cor}}^T \otimes GCN(\mathbf{F_{cor}} \otimes \mathbf{F}_\theta). \quad (7)$$

Finally, the reconstructed features $\mathbf{G}'$ are combined with the features $\mathbf{W}_\rho(\mathbf{F}_{AFS_3})$ to obtain the output $\mathbf{G}$:

$$\mathbf{G} = \mathbf{W}_\rho(\mathbf{F}_{AFS_3}) + \mathbf{W_z}(\mathbf{G}'), \quad (8)$$

where $\mathbf{W_z}$ denotes 1×1 convolution.

3.3 Discrepancy-Aggregation Detector (DisAD)

The double-head reconstructed forged images essentially differ from the input forged images in visual appearance [Cao *et al.*, 2022]. To further explore the probable forgery regions within reconstructed images, based on the DouHR module, we further introduce a DisAD head network via two **Reconstruction-guidance Feature Aggregation (RFA)** modules to aggregate the obtained genuine compact visual patterns (*i.e.*, $\hat{\mathbf{X}}_1$ of AFS and $\hat{\mathbf{X}}_2$ of SAM), resulting in an improved forgery detection capability on unknown patterns (*i.e.*, the greater generalization capability).

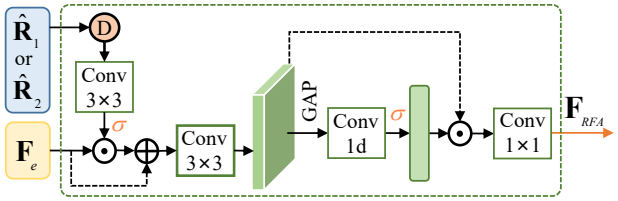

Figure 4: Illustration of Reconstruction-guidance Feature Aggregation (RFA) module, which can aggregate the obtained genuine compact visual patterns, such that the forgery detection capability on unknown patterns can be improved.

As shown in Figure 1, we first calculate the differences between two reconstructed images (*i.e.*, $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$) and the original input image $\mathbf{X}$ in discrepancy-aggregation detector. The pixel-level difference masks are expressed as:

$$\hat{\mathbf{R}}_1 = \left|\mathbf{X} - \hat{\mathbf{X}}_\mathbf{1}\right|, \hat{\mathbf{R}}_1 = \left|\mathbf{X} - \hat{\mathbf{X}}_\mathbf{2}\right|, \quad (9)$$

where $|\cdot|$ refers to the absolute value function. Then for RFA in Figure 4, given difference masks $\hat{\mathbf{R}}_1$ or $\hat{\mathbf{R}}_2$ and the summation of textural discrepancy information and encoding feature $\mathbf{F}_e = \mathbf{D}_4 \oplus \mathbf{F}_5$, we perform an element-wise multiplication between them with a residual connection and a 3×3 convolution to obtain the fused features $\mathbf{F}_d$:

$$\mathbf{F}_d = f_{c3}(\mathbf{F}_e \odot (\sigma(f_{c3}(D(\hat{R}_{1/2})))) \oplus \mathbf{F}_e), \quad (10)$$

where D denotes down-sampling and f_{c3} is 3×3 convolution. σ is a sigmoid function and $\oplus$ is element-wise addition. To enhance feature representations, inspired [Wang *et al.*, 2020], we aggregate the features $\mathbf{F}_d$ using a channel-wise global average pooling (GAP). Then the channel weight is obtained by the 1D convolution followed by a sigmoid function. Finally, the channel attention is multiplied with the input

features $\mathbf{F}_d$ to obtain the final output $\mathbf{F}_{RFA}$, *i.e.*,

$$\mathbf{F}_{RFA} = f_{c1}(\sigma(f_{1d}(GAP(\mathbf{F}_d))) \odot \mathbf{F}_d), \tag{11}$$

where f_{c1} is 1×1 convolution and f_{1d} is 1D convolution.

3.4 Loss Function

DisGRL has two kinds of supervision: the image-level binary classification label based on the cross-entropy loss (*i.e.*, L_{cls}), and the pixel-level reconstruction learning label. During training, we employ the reconstruction loss (L_{r1} and L_{r2}) [Cao *et al.*, 2022] between real images and their two reconstructed images. Besides, a metric-learning loss (*i.e.*, L_m) [Cao *et al.*, 2022] based on $\mathbf{F}_5$ is used to enhance the reconstruction difference to facilitate model learning. Thus, the total loss can be expressed as:

$$L_{total} = L_{cls} + \lambda_1 L_{r1} + \lambda_2 L_{r2} + \lambda_3 L_m, \tag{12}$$

where λ is a trade-off hyper-parameter for loss balance.

4 Experiments

4.1 Experimental Settings

Datasets. To facilitate a fair result comparison with state-of-the-art methods, we conducted experiments on four fundamental yet challenging face forgery datasets, including FaceForensics++ (FF++) [Rössler *et al.*, 2019], Celeb-DF [Li *et al.*, 2020b], WLD [Zi *et al.*, 2020], and DFDC [Dolhansky *et al.*, 2019]. Due to the page limit, details of each dataset are given in supplementary materials.

Implementation Details. We implemented our model on the PyTorch framework and used Xception [Chollet, 2017] pretrained on ImageNet [Deng *et al.*, 2009] as our mainstream backbone. The input face images are resized into 299×299 and augmented by random horizontal flipping. In the training phase, the batch size is set to 32, and Adam optimizer [Kingma and Ba, 2015] with learning rate 1e-4, and weight decay 1e-5 are adopted to optimize the model. The step learning rate strategy with a gamma of 0.5 is utilized to adjust the learning rate. Following [Cao *et al.*, 2022], λ_1, λ_2, and λ_3 in Eq. (12) are empirically set to 0.1.

Evaluation Metrics. In this work, we reported results on the commonly used evaluation metrics [Cao *et al.*, 2022; Sun *et al.*, 2022b; Zhuang *et al.*, 2022], including Accuracy (ACC), Area Under the Curve (AUC), and Equal Error Rate (EER).

4.2 Quantitative Results

To demonstrate the effectiveness of our proposed method, we compare it with the state-of-the-art methods, *i.e.*, Xception [Rössler *et al.*, 2019], Two-branch [Masi *et al.*, 2020], SPSL [Liu *et al.*, 2021], RFM [Wang and Deng, 2021], Freq-SCL [Li *et al.*, 2021], Add-Net [Zi *et al.*, 2020], F^3-Net [Qian *et al.*, 2020], MAT [Zhao *et al.*, 2021], RECCE [Cao *et al.*, 2022], ITA-SIA [Sun *et al.*, 2022a], Multi-task [Nguyen *et al.*, 2019], MLDG [Li *et al.*, 2018], LTW [Sun *et al.*, 2021], and DCL [Sun *et al.*, 2022b]. For a fair comparison, all experimental results of these methods which we employ for comparisons are either explicitly cited from works or generated by models that are retrained with open-source codes.

Intra-Dataset Evaluation. Table 1 shows result comparisons with our DisGRL against 10 competitors under intra-dataset evaluations. We can observe that DisGRL consistently outperforms other models on FF++ [Rössler *et al.*, 2019], WLD [Zi *et al.*, 2020], and DFDC [Dolhansky *et al.*, 2019]. Especially on the challenging WLD, our method still surpasses the second-best RECCE by 1.25% in terms of AUC. This suggests that when confronted with more identities from real-world scenes, our method owns the superior ability to detect discrepancies between real faces and fake ones. On Celeb-DF [Li *et al.*, 2020b], though ITA-SIA achieves the highest AUC, our DisGRL still achieves comparable results on the other datasets, especially on the low-quality setting of the FF++ ($\uparrow$ 1.74%). Different from ITA-SIA which introduces a self-information metric to enhance the feature representation, DisGRL produces a more robust representation through double-head reconstruction, which works well in conjunction with single reconstruction for forgery detection.

Cross-Dataset Evaluation. To explore the generalization of our method on unseen datasets compared with recent general face forgery detection methods, we focus on the more challenging cross-dataset evaluation. Table 2 reports the quantitative results by training the models on FF++ (LQ) [Rössler *et al.*, 2019] and testing them on Celeb-DF [Li *et al.*, 2020b], WLD [Zi *et al.*, 2020], and DFDC [Dolhansky *et al.*, 2019], accordingly. It can be concluded that our method achieves a certain improvement in generalization ability by taking good advantage of double-head reconstruction structures. In particular, the AUC score of our method on Celeb-DF ($\uparrow$ 1.32%), WLD ($\uparrow$ 2.42%), and DFDC ($\uparrow$ 1.83%) datasets is enhanced when compared with RECCE. Overall, our method promotes the extraction of genuine compact visual patterns and can be generalized to unseen forgeries rather than modeling the pattern of the single forgery techniques.

Cross-Manipulation Evaluation. To further demonstrate the generalization among different manipulated manners, we conduct the fine-grained cross-manipulation evaluation by training a model on one specific method and testing it on all four methods listed in FF++ (LQ). As shown in Table 3, our DisGRL generally outperforms the competitors in most cases, including both intra-manipulation (diagonal of the table) results and cross-manipulation. Specifically, when training on NT and testing on F2F, though MAT is equipped with EfficientNet-b4, our DisGRL based on Xception still outperforms it by a margin of 2.69%. Additionally, a 2.41% performance gain in terms of AUC is achieved by our method compared with RECCE, which illustrates that it is feasible to explore common features of real faces to distinguish real and fake faces. With help of the double-head reconstruction strategy and carefully designed cascaded discrepancy external attention, our method exceeds all other methods in terms of the average AUC of cross-manipulation evaluations.

Multi-Source Manipulation Evaluation. Multi-source manipulation evaluation refers to situations in which the forged techniques utilized for training are not restricted to just one way. Following the LTW [Sun *et al.*, 2021] and DCL [Sun *et al.*, 2022b], we conduct experiments on the low-quality (LQ) version of FF++ [Rössler *et al.*, 2019] to demonstrate the practicality of our method in real-world scenarios. As

Methods	Pub./Year	FF++ HQ		FF++ LQ		Celeb-DF		WLD		DFDC	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Xception	ICCV'19	95.73	96.30	86.86	89.30	97.90	99.73	77.25	86.76	79.35	89.50
Two branch	ECCV'20	96.43	98.70	86.34	86.59	–	–	–	–	–	–
Add-Net	ACM MM'20	96.78	97.74	87.50	91.01	96.93	99.55	76.25	86.17	78.71	89.85
F^3-Net	ECCV'20	97.52	98.10	90.43	93.30	95.95	98.93	80.66	87.53	76.17	88.39
SPSL	CVPR'21	91.50	95.32	81.57	82.82	–	–	–	–	–	–
RFM	CVPR'21	95.69	98.79	87.06	89.83	97.96	99.94	77.38	83.92	80.83	89.75
Freq-SCL	CVPR'21	96.69	99.28	89.00	92.39	–	–	–	–	–	–
MAT	CVPR'21	97.60	99.29	88.69	90.40	97.92	99.94	82.86	90.71	76.81	90.32
RECCE	CVPR'22	97.06	99.32	91.03	95.02	98.59	99.94	83.25	92.02	81.20	91.33
ITA-SIA	ECCV'22	97.64	99.35	90.23	93.45	98.48	99.96	83.95	91.34	–	–
DisGRL	IJCAI'23	97.69	99.48	91.27	95.19	98.71	99.91	84.53	93.27	82.35	92.50

Table 1: Intra-dataset evaluation and result comparisons on four benchmarks. "HQ" and "LQ" denote the High-Quality version and the Low-Quality version of the corresponding dataset, respectively. The top three results are highlighted in red, green, and blue, respectively.

Methods	Celeb-DF		WLD		DFDC	
	AUC	EER	AUC	EER	AUC	EER
Xception	61.80	41.73	62.72	40.65	63.61	40.58
F^3-Net	61.51	42.03	57.10	45.12	64.60	39.84
Add-Net	65.29	38.90	62.35	41.42	64.78	40.23
RFM	65.63	38.54	57.75	45.45	66.01	39.05
MAT	67.02	37.90	59.74	43.73	68.01	37.17
RECCE	68.71	35.73	64.31	40.53	69.06	36.08
DisGRL	**70.03**	**34.23**	**66.73**	**39.24**	**70.89**	**34.27**

Table 2: Cross-dataset result evaluation on FF++ (LQ), Celeb-DF, WLD, and DFDC in terms of AUC ↑ (%) and EER ↓ (%).

shown in Table 4, we can observe that our DisGRL obtains cutting-edge performance in terms of AUC and ACC on all protocols. In particular, DisGRL outperforms the recent DCL by around 7% in the setting of GID-F2F, proving its durability and ability to ensure generalization under various scenarios.

4.3 Ablation Study

To validate the effectiveness of each component, we designed several ablation experiments on the WildDeepfake dataset in varied configurations with the components added progressively. As shown in Table 5, the setup model variants are as follows: for the baseline model of a), we follow the classic image classification pipeline, *i.e.*, Xception [Chollet, 2017]. b) and c) the encoder-decoder backbone with the introduction of a single-head reconstruction (Rec-1 or Rec-2) learning scheme. d), the encoder-decoder backbone equipped with double-head reconstruction (Rec-1 and Rec-2). For the model of e), we remove the RFA and adopt the element-wise addition to replace it, f) is the proposed DisGRL without DEA, and g) is our DisGRL.

Effectiveness of DouHR. We can observe in Table 5 that the double-head reconstruction learning module performs better than the baseline model of a) and its variant b), c) Baseline + single-head. Therefore, Rec-2, as the complementary information of Rec-1, aims to capture the genuine compact visual pattern of real regions and fake regions, which is beneficial to boost detection performance.

Methods	Train	DF	F2F	FS	NT	CAvg.
Freq-SCL	DF	98.91	58.90	66.87	63.61	63.13
MAT		99.51	66.41	67.33	66.01	66.58
RECCE		99.65	70.66	74.29	67.34	70.76
DisGRL		**99.67**	**71.76**	**75.21**	**68.74**	**71.90**
Freq-SCL	F2F	67.55	93.06	55.35	66.66	63.19
MAT		73.04	97.96	65.10	71.88	70.01
RECCE		**75.99**	98.06	64.53	72.32	70.95
DisGRL		75.73	**98.69**	**65.71**	**74.15**	**71.86**
Freq-SCL	FS	75.90	54.64	98.37	49.72	60.09
MAT		82.33	61.65	98.82	54.79	66.26
RECCE		82.39	64.44	98.82	56.70	67.84
DisGRL		**82.73**	**64.85**	**99.01**	**56.96**	**68.18**
Freq-SCL	NT	79.09	74.21	53.99	88.54	69.10
MAT		74.56	80.61	60.90	93.34	72.02
RECCE		78.83	80.89	63.70	93.63	74.47
DisGRL		**80.29**	**83.30**	**65.23**	**94.10**	**76.27**

Table 3: Cross-manipulation evaluation in terms of AUC (%). Diagonal results indicate intra-domain performance. DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT) are four image manipulation approaches in FF++ [Rössler *et al.*, 2019]. "CAvg." denotes the average of cross-manipulation evaluations.

Effectiveness of DisGE. Then the model of e) DisGRL w/o RFA achieves better overall performance compared with the model of d) Baseline + double-head, especially in terms of AUC with 1.37 % performance gains. It verifies that DEA enhances the model's efficiency to mine the forgery-sensitive visual pattern within the instance by cascading shallow and deep features in the encoder to focus on image forgery cues rather than on semantic image content. Therefore, it is feasible to improve the classification learning capabilities of the detector when combined with the integrated representation collected by the decoder, leading to a larger performance increase for variation e) DisGRL w/o RFA.

Effectiveness of DisAD. The comparison between variants d) Baseline + double-head and f) DisGRL w/o DEA in Table 5 can demonstrate the effectiveness of our proposed RFA, which aggregates the obtained genuine compact visual pat-

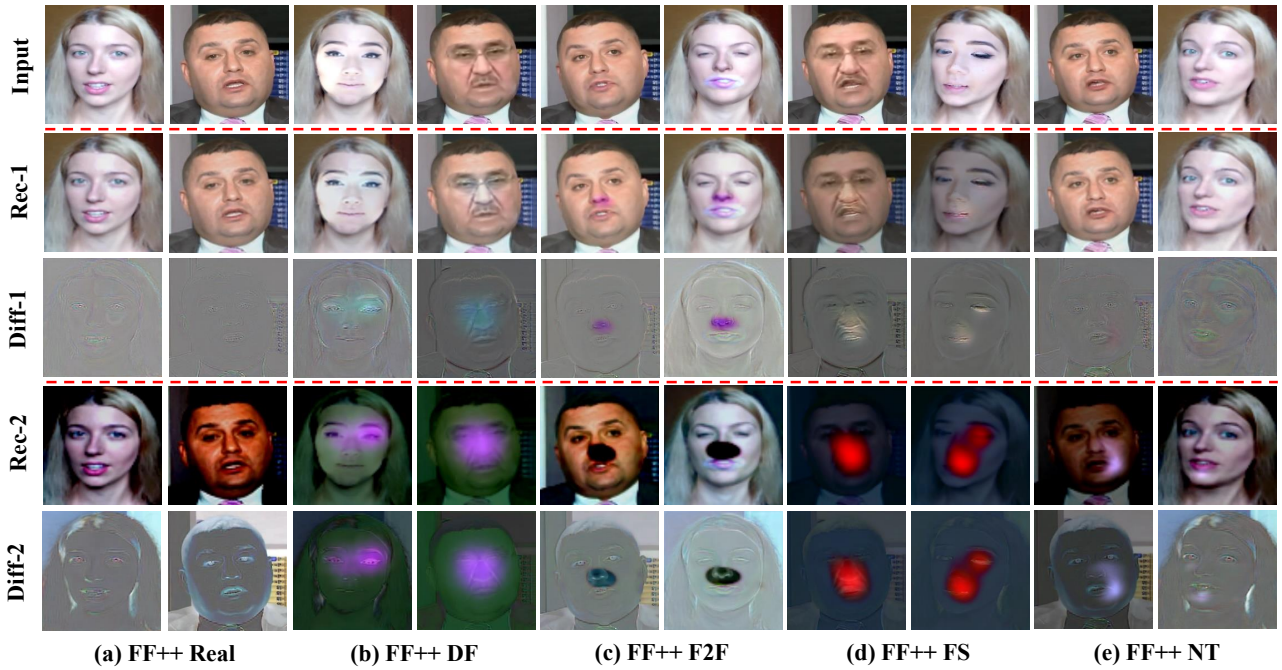

Figure 5: Reconstruction and differential visualization of the proposed model on the FaceForensics++ dataset. "Rec-1" and "Rec-2" are the double reconstructions results. "Diff-1" and "Diff-2" denote the corresponding pixel-level difference, respectively.

Methods	GID-DF	GID-F2F	GID-FS	GID-NT
Multi-task	66.8/–	56.5/–	51.7/–	56.0/–
MLDG	67.2/73.1	58.1/61.7	58.1/61.7	56.9/60.7
LTW	69.1/75.6	65.7/72.4	62.5/68.1	58.5/60.8
DCL	75.9/83.8	67.9/75.1	–/–	–/–
DisGRL	**77.3/86.1**	**75.8/84.3**	**76.9/86.3**	**66.3/72.8**

Table 4: Multi-source results on ACC (%)/AUC (%).

NO	B	Rec-1	Rec-2	DEA	RFA	ACC	AUC
a)	✓					77.13	86.21
b)	✓	✓				80.76	88.47
c)	✓		✓			81.37	88.94
d)	✓	✓	✓			82.84	90.98
e)	✓	✓	✓	✓		83.68	92.35
f)	✓	✓	✓		✓	83.24	91.68
g)	✓	✓	✓	✓	✓	**84.53**	**93.27**

Table 5: Ablation studies on WildDeepfake [Zi *et al.*, 2020] in terms of ACC (%) and AUC (%).

tern and emphasizes the probably forged regions. And combining all the proposed components can achieve the best performance in terms of ACC and AUC scores.

4.4 Visualizations

Our proposed reconstruction learning aims to preserve more variations by building a double-head reconstruction scheme. To validate its effectiveness, as illustrated in Figure 5, we visualize the outputs of the two reconstructions and the corresponding difference masks between the original input and reconstruction maps. We can observe that the real faces can be well reconstructed with little blurring, while the forged portions of the fake ones cannot be recovered. Difference masks, indicating possible traces of forged areas, further amplify the differences between real and forged faces. Compared with Diff-1, Diff-2 is able to additionally enhance and complement the forged areas in faces. For instance, NT operates around the mouth region and the response in Diff-1 of the corresponding sample is weak around the mouth region, while the value is larger in Diff-2, illustrating the importance and usefulness of an additional head for reconstruction in image forgery detection.

5 Conclusion

In this work, we proposed a novel image forgery detection paradigm, termed DisGRL, to improve the model learning capacity on forgery-sensitive and genuine compact visual patterns. DisGRL primarily consisted of a discrepancy-guided encoder, a decoder, a double-head reconstruction module, and a discrepancy-aggregation detector head network. The advantage of DisGRL was that it can not only encode general semantic features but also enhance the forgery cues of the given image. Experimental results on four widely used face forgery datasets validated the effectiveness of our proposed method against state-of-the-art competitors on both seen and unseen forgeries. DisGRL is a general paradigm, which can be used in general image forgery detection tasks. Therefore, in the future, we will explore how to apply DisGRL to more challenging natural scene datasets in terms of quantity and quality. Besides, exploring how to use DisGRL in the forgery detection of video data is also a promising research direction.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62276112), the National Natural Science Foundation of China Regional Joint Fund of NSFC (U19A2057), Jilin Province Science and Technology Development Plan Key R&D Project (20230201088GX), and Collaborative Innovation Project of Anhui Universities (GXXT-2022-044).

References

[Cao *et al.*, 2022] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, pages 4103–4112, 2022.

[Chen *et al.*, 2021a] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *AAAI*, pages 1081–1088, 2021.

[Chen *et al.*, 2021b] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *ICCV*, pages 14165–14173, 2021.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1800–1807, 2017.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Dolhansky *et al.*, 2019] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton-Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[Dong *et al.*, 2021] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[Fei *et al.*, 2022] Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, and Jian Weng. Learning second order local anomaly for general face forgery detection. In *CVPR*, pages 20238–20248, 2022.

[Gu *et al.*, 2022] Zhihao Gu, Taiping Yao, Yang Chen, Ran Yi, Shouhong Ding, and Lizhuang Ma. Region-aware temporal inconsistency learning for deepfake video detection. In *IJCAI*, pages 920–926, 2022.

[Guo *et al.*, 2023] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5436–5447, 2023.

[Haliassos *et al.*, 2021] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, pages 5039–5049, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hu *et al.*, 2021] Ziheng Hu, Hongtao Xie, Yuxin Wang, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Dynamic inconsistency-aware deepfake video detection. In *IJCAI*, pages 736–742, 2021.

[Jia *et al.*, 2022] Shuai Jia, Chao Ma, Taiping Yao, Bangjie Yin, Shouhong Ding, and Xiaokang Yang. Exploring frequency adversarial attacks for face forgery detection. In *CVPR*, pages 4093–4102, 2022.

[Jiang *et al.*, 2020] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, pages 2886–2895, 2020.

[Kawar *et al.*, 2022] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, pages 1–15, 2015.

[Li *et al.*, 2018] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, pages 3490–3497, 2018.

[Li *et al.*, 2020a] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, pages 5000–5009, 2020.

[Li *et al.*, 2020b] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*, pages 3204–3213, 2020.

[Li *et al.*, 2021] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *CVPR*, pages 6458–6467, 2021.

[Li *et al.*, 2022a] Jiaming Li, Hongtao Xie, Lingyun Yu, and Yongdong Zhang. Wavelet-enhanced weakly supervised local feature learning for face forgery detection. In *ACM MM*, pages 1299–1308, 2022.

[Li *et al.*, 2022b] Lei Li, Kai Fan, and Chun Yuan. Cross-modal representation learning and relation reasoning for bidirectional adaptive manipulation. In *IJCAI*, pages 3222–3228, 2022.

[Liu *et al.*, 2021] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *CVPR*, pages 772–781, 2021.

[Lu *et al.*, 2019] Yi Lu, Yaran Chen, Dongbin Zhao, and Jianxin Chen. Graph-fcn for image semantic segmentation. In *ISNN*, volume 11554, pages 97–105, 2019.

[Masi *et al.*, 2020] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, volume 12352, pages 667–684, 2020.

[Nguyen *et al.*, 2019] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *BTAS*, pages 1–8, 2019.

[Qian *et al.*, 2020] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, volume 12357, pages 86–103, 2020.

[Rao and Ni, 2021] Yuan Rao and Jiangqun Ni. Self-supervised domain adaptation for forgery localization of JPEG compressed images. In *ICCV*, pages 15014–15023, 2021.

[Rao *et al.*, 2022] Yuan Rao, Jiangqun Ni, Weizhe Zhang, and Jiwu Huang. Towards jpeg-resistant image forgery detection and localization via self-supervised domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–12, 2022.

[Robert *et al.*, 2018] Thomas Robert, Nicolas Thome, and Matthieu Cord. Hybridnet: Classification and reconstruction cooperation for semi-supervised learning. In *ECCV*, volume 11211, pages 158–175, 2018.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022.

[Rössler *et al.*, 2019] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019.

[Shi *et al.*, 2023] Zenan Shi, Haipeng Chen, and Dong Zhang. Transformer-auxiliary neural networks for image manipulation localization by operator inductions. *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[Sun *et al.*, 2021] Ke Sun, Hong Liu, Qixiang Ye, Yue Gao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. Domain general face forgery detection by learning to weight. In *AAAI*, pages 2638–2646, 2021.

[Sun *et al.*, 2022a] Ke Sun, Hong Liu, Taiping Yao, Xiaoshuai Sun, Shen Chen, Shouhong Ding, and Rongrong Ji. An information theoretic approach for attention-driven face forgery detection. In *ECCV*, volume 13674, pages 111–127, 2022.

[Sun *et al.*, 2022b] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *AAAI*, pages 2316–2324, 2022.

[Wang and Deng, 2021] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *CVPR*, pages 14923–14932, 2021.

[Wang *et al.*, 2020] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, pages 11531–11539, 2020.

[Wang *et al.*, 2022a] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *CVPR*, pages 2354–2363, 2022.

[Wang *et al.*, 2022b] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17662–17672, 2022.

[Yan *et al.*, 2023] Rui Yan, Lingxi Xie, Xiangbo Shu, Liyan Zhang, and Jinhui Tang. Progressive instance-aware feature learning for compositional action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–14, 2023.

[Yoshihashi *et al.*, 2019] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *CVPR*, pages 4016–4025, 2019.

[Zhang *et al.*, 2020] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *ECCV*, volume 12373, pages 323–339, 2020.

[Zhang *et al.*, 2021] Daichi Zhang, Chenyu Li, Fanzhao Lin, Dan Zeng, and Shiming Ge. Detecting deepfake videos with temporal dropout 3dcnn. In *IJCAI*, pages 1288–1294, 2021.

[Zhang *et al.*, 2022a] Dong Zhang, Yi Lin, Hao Chen, Zhuotao Tian, Xin Yang, Jinhui Tang, and Kwang Ting Cheng. Deep learning for medical image segmentation: tricks, challenges and future directions. *arXiv preprint arXiv:2209.10307*, 2022.

[Zhang *et al.*, 2022b] Dong Zhang, Jinhui Tang, and Kwang-Ting Cheng. Graph reasoning transformer for image parsing. In *ACM MM*, pages 2380–2389, 2022.

[Zhao *et al.*, 2021] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, pages 2185–2194, 2021.

[Zhu *et al.*, 2020] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain GAN inversion for real image editing. In *ECCV*, volume 12362, pages 592–608, 2020.

[Zhuang *et al.*, 2022] Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *ECCV*, volume 13665, pages 391–407, 2022.

[Zi *et al.*, 2020] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *ACM MM*, pages 2382–2390, 2020.