Discrepancy-Guided Reconstruction Learning for Image Forgery Detection

Zenan Shi\textsuperscript{1,2}, Haipeng Chen\textsuperscript{1,2}, Long Chen\textsuperscript{3}, Dong Zhang\textsuperscript{3,*}

\textsuperscript{1}College of Computer Science and Technology, Jilin University
\textsuperscript{2}Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University
\textsuperscript{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 accessorional 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.

*Corresponding author: Dong Zhang.
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 $F_1$ to connect feature maps from the corresponding encoder network layer, which finally has the same scale as $F_2$. 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; Haliasoss 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 $X$, 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., $F_i \quad (i = 1, 2, \ldots, 5)$, and the Discrepancy External Attention (DEA) branch is applied to different level feature to extract explicit discrepancy-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 $D_i$ is expressed as:

$$D_i = \begin{cases} \text{Dea}(F_i), & i = 1 \\ \text{Dea}(\text{Cat}(D_{i-1}, F_i)), & i \in [2, 3, 4] \end{cases}$$

(1)

where $\text{Dea}(\cdot)$ and $\text{Cat}(\cdot)$ 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 $3 \times 3$ convolutional layer on the input feature maps $F \in \mathbb{R}^{C \times H \times W}$ with the same channel size $C$. Then, an adaptive average pooling is used to obtain the pooled features $F_d$. After that, the differentiated maps can be obtained through $D' = F - F_d$ 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 $D' \in \mathbb{R}^{C \times H \times W}$, 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 $D$ can be obtained through a $1 \times 1$ 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.](image)

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., 2022; 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 $\hat{X}_1$ 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 $\hat{X}_2$. The DouHR module can be formulated as:

$$\hat{X}_1 = \text{AFS}(F_{AFS_1}, F_1), \quad \hat{X}_2 = \text{SAM}(F_{AFS_1}, F_1),$$

(2)

where $F_{AFS_1}$ 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 $1 \times 1$ 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 $F_{AFS_2}$ and $F_2$. In DouHR, AFS receives the $F_{AFS_1}$ and as $F_1$ input. The concatenation operation $\text{Cat}(\cdot)$ is first carried out on $F_{AFS_1}$ and $F_1$ in the channel dimension, i.e., $\tilde{F} = \text{Cat}(F_{AFS_1}, F_1)$, followed by a depthwise separable convolution $f_{d3}$ to obtain attention map $A_{att}$ 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:

$$A_{att} = \sigma(f_{d3}(\tilde{F})), \quad A = f_{d3}(f_{d3}(\tilde{F}) \circ A_{att}) + f_{d3}(\tilde{F}),$$

(3)

where $f_{d3}$ and $\sigma(\cdot)$ are the $3 \times 3$ convolution layer and sigmoid activation function, respectively. Other AFS procedures are calculated in a similar way.
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 $F_{AFS}$, we first apply three $1 \times 1$ convolutions (i.e., $W_{\rho}, W_{\theta},$ and $W_{\varphi}$) to reduce the channel dimension into 16, and obtain feature maps $F_{\theta}, F_{\varphi}$, which can be expressed as:

$$F_{\theta} = W_{\theta}(W_{\rho}(F_{AFS})), \quad F_{\varphi} = W_{\varphi}(W_{\rho}(F_{AFS})). \quad (4)$$

For $F_{1}$, we down-sample it to the same size as $W_{\rho}(F_{AFS})$. Then we apply a Softmax function along the channel dimension and calculate the element-wise multiplication with $F_{\varphi}$ 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:

$$F_{w} = AVP(F_{\varphi} \odot \text{softmax}(D(F_{1}))), \quad (5)$$

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

$$F_{cor} = \text{softmax}(F_{w} \odot (F_{\varphi})^{T}). \quad (6)$$

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

$$G' = F_{cor}^{T} \odot \text{GCN}(F_{cor} \odot F_{\theta}). \quad (7)$$

Finally, the reconstructed features $G'$ are combined with the features $W_{\rho}(F_{AFS})$ to obtain the output $G$:

$$G = W_{\rho}(F_{AFS}) + W_{z}(G'), \quad (8)$$

where $W_{z}$ denotes $1 \times 1$ 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., $\hat{X}_{1}$ of AFS and $\hat{X}_{2}$ of SAM), resulting in an improved forgery detection capability on unknown patterns (i.e., the greater generalization capability).

As shown in Figure 1, we first calculate the differences between two reconstructed images (i.e., $\hat{X}_{1}$ and $\hat{X}_{2}$) and the original input image $X$ in discrepancy-aggregation detector. The pixel-level difference masks are expressed as:

$$\tilde{R}_{1} = |X - \hat{X}_{1}|, \quad \tilde{R}_{2} = |X - \hat{X}_{2}|, \quad (9)$$

where $|\cdot|$ refers to the absolute value function. Then for RFA in Figure 4, given difference masks $\tilde{R}_{1}$ or $\tilde{R}_{2}$ and the summation of textural discrepancy information and encoding feature $F_{e} = D_{4} \oplus F_{5}$, we perform an element-wise multiplication between them with a residual connection and a $3 \times 3$ convolution to obtain the fused features $F_{d}$:

$$F_{d} = f_{33}(F_{e} \odot (\sigma(f_{33}(D_{1}(\hat{R}_{1}/2)))) \oplus F_{e}), \quad (10)$$

where $D$ denotes down-sampling and $f_{33}$ is $3 \times 3$ convolution. $\sigma$ is a sigmoid function and $\oplus$ is element-wise addition. To enhance feature representations, inspired [Wang et al., 2020], we aggregate the features $F_{d}$ 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 $F_d$ to obtain the final output $F_{RFA}$, i.e.,
\[
F_{RFA} = f_{c1}(\sigma(f_{id}(GAP(F_d))) \odot F_d),
\]
where $f_{c1}$ is $1 \times 1$ convolution and $f_{id}$ 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 ($L_{r_1}$ and $L_{r_2}$) [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 $F_5$ is used to enhance the reconstruction difference to facilitate model learning. Thus, the total loss can be expressed as:
\[
L_{total} = L_{cls} + \lambda_1 L_{r_1} + \lambda_2 L_{r_2} + \lambda_3 L_m,
\]
where $\lambda$ 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] pre-trained on ImageNet [Deng et al., 2009] as our mainstream backbone. The input face images are resized into $299 \times 299$ 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], $\lambda_1$, $\lambda_2$, and $\lambda_3$ 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], FreqSCL [Li et al., 2021], Add-Net [Zi et al., 2020], F$^1$-Net [Qian et al., 2020], MAT [Zhao et al., 2021], RECCCE [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 in-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 when 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
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].

<table>
<thead>
<tr>
<th>Methods</th>
<th>FF++ HQ</th>
<th>FF++ LQ</th>
<th>Celeb-DF</th>
<th>WLD</th>
<th>DFDC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC</td>
<td>AUC</td>
<td>ACC</td>
<td>AUC</td>
<td>ACC</td>
</tr>
<tr>
<td>Xception</td>
<td>95.73</td>
<td>96.30</td>
<td>86.76</td>
<td>89.30</td>
<td>97.90</td>
</tr>
<tr>
<td>Two branch</td>
<td>96.43</td>
<td>98.70</td>
<td>86.34</td>
<td>86.59</td>
<td>–</td>
</tr>
<tr>
<td>Add-Net</td>
<td>ACM MM’20</td>
<td>96.78</td>
<td>97.74</td>
<td>87.50</td>
<td>91.01</td>
</tr>
<tr>
<td>F³-Net</td>
<td>ECCV’20</td>
<td>97.52</td>
<td>98.10</td>
<td>90.43</td>
<td>93.30</td>
</tr>
<tr>
<td>SPSL</td>
<td>CVPR’21</td>
<td>91.50</td>
<td>95.32</td>
<td>81.57</td>
<td>82.82</td>
</tr>
<tr>
<td>RFM</td>
<td>CVPR’21</td>
<td>95.69</td>
<td>98.79</td>
<td>87.06</td>
<td>89.83</td>
</tr>
<tr>
<td>Freq-SCL</td>
<td>CVPR’21</td>
<td>96.69</td>
<td>99.28</td>
<td>89.00</td>
<td>92.39</td>
</tr>
<tr>
<td>MAT</td>
<td>CVPR’21</td>
<td>97.60</td>
<td>99.29</td>
<td>88.69</td>
<td>90.40</td>
</tr>
<tr>
<td>RECCCE</td>
<td>CVPR’22</td>
<td>97.06</td>
<td>99.32</td>
<td>91.03</td>
<td>95.02</td>
</tr>
<tr>
<td>ITA-SIA</td>
<td>ECCV’22</td>
<td>97.64</td>
<td>99.35</td>
<td>90.23</td>
<td>93.45</td>
</tr>
<tr>
<td>DisGRL</td>
<td>IJCAI’23</td>
<td>97.69</td>
<td>99.48</td>
<td>91.27</td>
<td>95.19</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Train</th>
<th>DF</th>
<th>F2F</th>
<th>FS</th>
<th>NT</th>
<th>CAvg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq-SCL</td>
<td>75.65</td>
<td>93.06</td>
<td>55.35</td>
<td>66.66</td>
<td>70.76</td>
<td></td>
</tr>
<tr>
<td>MAT</td>
<td>97.34</td>
<td>99.46</td>
<td>65.10</td>
<td>71.88</td>
<td>70.01</td>
<td></td>
</tr>
<tr>
<td>RECCCE</td>
<td>75.99</td>
<td>98.06</td>
<td>64.53</td>
<td>72.32</td>
<td>70.95</td>
<td></td>
</tr>
<tr>
<td>DisGRL</td>
<td>75.73</td>
<td>98.69</td>
<td>65.71</td>
<td>74.15</td>
<td>71.86</td>
<td></td>
</tr>
</tbody>
</table>

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

### 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.

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

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

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

<table>
<thead>
<tr>
<th>NO</th>
<th>B</th>
<th>Rec-1</th>
<th>Rec-2</th>
<th>DEA</th>
<th>RFA</th>
<th>ACC</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>a)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>77.13</td>
<td>86.21</td>
</tr>
<tr>
<td>b)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>80.76</td>
<td>88.47</td>
</tr>
<tr>
<td>c)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>81.37</td>
<td>88.94</td>
</tr>
<tr>
<td>d)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>82.84</td>
<td>90.98</td>
</tr>
<tr>
<td>e)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>83.68</td>
<td>92.35</td>
</tr>
<tr>
<td>f)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>83.24</td>
<td>91.68</td>
</tr>
<tr>
<td>g)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>84.53</td>
<td>93.27</td>
</tr>
</tbody>
</table>

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


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


