# End-to-End Reconstruction-Classification Learning for Face Forgery Detection

Junyi Cao[1] Chao Ma[1*] Taiping Yao[2] Shen Chen[2] Shouhong Ding[2] Xiaokang Yang[1]
[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2] Youtu Lab, Tencent
{junyicao, chaoma, xkyang}@sjtu.edu.cn {taipingyao, kobeschen, ericshding}@tencent.com

## Abstract

*Existing face forgery detectors mainly focus on specific forgery patterns like noise characteristics, local textures, or frequency statistics for forgery detection. This causes specialization of learned representations to known forgery patterns presented in the training set, and makes it difficult to detect forgeries with unknown patterns. In this paper, from a new perspective, we propose a forgery detection framework emphasizing the common compact representations of genuine faces based on reconstruction-classification learning. Reconstruction learning over real images enhances the learned representations to be aware of forgery patterns that are even unknown, while classification learning takes the charge of mining the essential discrepancy between real and fake images, facilitating the understanding of forgeries. To achieve better representations, instead of only using the encoder in reconstruction learning, we build bipartite graphs over the encoder and decoder features in a multi-scale fashion. We further exploit the reconstruction difference as guidance of forgery traces on the graph output as the final representation, which is fed into the classifier for forgery detection. The reconstruction and classification learning is optimized end-to-end. Extensive experiments on large-scale benchmark datasets demonstrate the superiority of the proposed method over state of the arts.*

## 1. Introduction

The recent years have witnessed the considerable progress of face forgery generation methods [2, 4, 11, 18, 20, 41, 42, 50, 53]. Owing to the success of deep learning, generating ultra-realistic fake facial images or videos is really easy. An attacker can take advantage of these techniques to make fake news, defame celebrities, or break authentication, leading to serious political, social, and security consequences [27]. To mitigate malicious abuse of face forgery, it is urgent to develop effective detection methods.

* Corresponding author.

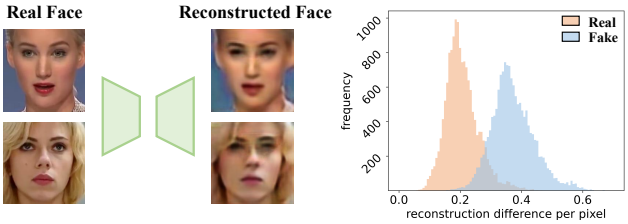


Figure 1. We perform reconstruction learning over only genuine samples to learn the common compact representations of real facial images (left). With the learned representations, the reconstruction difference of real and fake faces significantly differs in distribution (right), which facilitates forgery detection.

Early face forgery detection methods [1, 8, 23, 31, 35, 60] typically follow the classic pipeline of learning convolutional neural networks (CNN) for image classification. With off-the-shelf CNN backbones, these methods directly take a facial image as input and then classify it as real or fake. However, these vanilla CNNs tend to seek forgeries on a limited region of faces, indicating that the detectors are short of the understanding of forgery [45]. Recent works resort to specific forgery patterns such as noise characteristics [12, 58], local textures [6, 14, 55], and frequency information [22, 33] to better detect forgery artifacts that resided in fake faces. Despite the demonstrated promising results, they always rely on forgery patterns that are possessed by a certain manipulation technique presented in the training set. Thus, in the real-world scenario, due to the emergence of new manipulation techniques and various types of perturbations, forgeries with unknown patterns easily cause existing methods to fail.

To address the above issues, we have two major considerations to enhance the learned representations for face forgery detection. First, for learning representations that can generalize to unknown forgery patterns, exploring the common characteristics of *genuine* faces is more suitable than overfitting to specific forgery patterns presented in the training set. As previous studies [5, 36] indicate that real samples possess a relatively compact distribution, the

compact representations learned with real images are more likely to distinguish unknown forgery patterns from genuine faces. Second, to ensure that the learned representations capture the essential discrepancy between real and fake images, it is desirable to enhance the network reasoning about forgery cues. As such, classification learning provides a better understanding of forgeries from a global perspective.

With the above considerations in mind, in this paper, we present a novel reconstruction-classification learning (RECCE) framework to detect face forgeries. The key idea of which is illustrated in Figure 1. For reconstruction learning, we propose a reconstruction network, which consists of an encoder and a decoder, to model the distributions of only real faces. Besides the reconstruction loss, we apply a metric-learning loss on the decoder to make real images close, while real and fake images far away, in the embedding space. This ensures that fake images with unknown forgery patterns are more likely to be recognized due to the learned distributional discrepancy.

Based on the above constraints, the discrepancy information which reveals forgery cues is progressively strengthened at the decoder side, finally resulting in sound reconstruction for real faces and poor reconstruction for fake images. Thus, to achieve complete representations, instead of using merely the encoder output as features, we also consider the decoder features. Inspired by the recent advances in graph modeling [44, 47, 56] which can model the feature relationship flexibly and adaptively, we build bipartite graphs over the encoder and decoder features to reason about forgery cues captured by decoder features. Since different face forgery techniques result in forged traces across various scales, we adopt the multi-scale mechanism during the reasoning process to mine the forgery clues comprehensively. Furthermore, in view of that the reconstruction difference indicates probably forged regions, we use the reconstruction difference as guidance to attend to the graph output as the final representations for classification learning. The reconstruction and classification learning are jointly optimized in an end-to-end manner.

In brief, the main contributions are as follows:

- From a new perspective, we propose the RECCE framework for face forgery detection which mines the common features of genuine faces. This enhances the learned representations able to separate fake faces even with unknown forgery patterns from real images.
- We build bipartite graphs over the encoder output and decoder features in a multi-scale fashion to help the network reason about forgery clues and exploit the reconstruction difference as guidance to attend to probably forged traces.
- Extensive experiments on benchmark datasets, including FaceForensics++ [35], WildDeepfake [60], and DFDC [9], validate the superiority of the proposed method over the state-of-the-art approaches.

## 2. Related Work

**Face Forgery Detection.** Many efforts have been made to improve the performance of face forgery detection [1, 13, 23, 29, 30, 40, 45, 46, 60]. Early works like [31, 35] use state-of-the-art image classification backbones, *e.g.*, VGGNet [39] and XceptionNet [7], to extract features from cropped facial images and perform binary classification. However, the CNN backbones inherited from image classification models emphasize category-level differences rather than the nuances between real and fake images. Recently, in view of that forged faces become more visually realistic, a number of works propose to further mine specific forgery patterns such as noise statistics, local textures, and frequency information to distinguish fake faces from real ones. For example, Zhou *et al*. [58] design a two-stream neural network, in which one branch uses the visual appearance and the other branch focuses on local noise patterns to detect forged faces. Zhao *et al*. [55] devise a multi-attentional face forgery detector that aggregates the texture features and high-level semantic features of multiple local parts to classify real and fake samples. Qian *et al*. [33] and Li *et al*. [22] take the frequency details into account and propose frequency-aware models to separate bonafide faces and forged faces. Despite the improved performance, these approaches mainly rely on the learned forgery patterns presented in the training samples, and thus they will experience an obvious performance drop when dealing with unknown forgery patterns.

**Reconstruction Learning.** Reconstruction learning has been wildly used for representation learning in unsupervised settings [16, 26, 32, 49, 51, 52]. It encourages the model to encode more information about the input so as to restore the input effectively. The objective of it is to model the distribution of input data in the embedding space [28, 34, 54]. Some prior works have explored reconstruction learning for face forgery detection. For instance, Nguyen *et al*. [30] use a reconstruction network but with a focus on multi-task learning. In [10], Du *et al*. propose a locality-aware autoencoder and use a pixel-wise mask to attend to the forgery region. Note that, these approaches perform reconstruction learning over all the face images regardless of real and fake samples. Hence the learned representations are not ensured to be generalized. Differently, our approach explicitly learns to model the distribution of genuine faces. This helps the learned representations are more likely to detect unknown forgery patterns due to distributional discrepancy between genuine faces and manipulated samples. Together with classification learning, the proposed end-to-end reconstruction-classification learning approach has shown superiority on large-scale benchmarks over state of the arts.

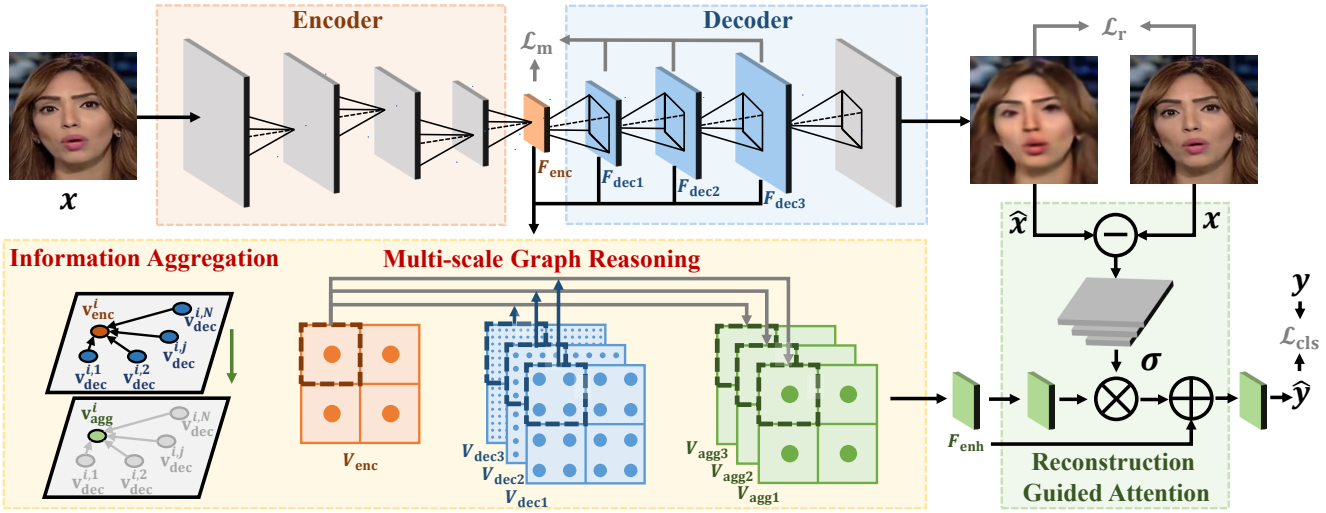


Figure 2. Schematic diagram of the proposed framework. The input images (real or fake faces) first enter an encoder-decoder reconstruction network for representation learning. The encoder output goes through the multi-scale graph reasoning module to achieve better representations, which are further guided by the reconstruction difference for final classification. The whole system is trained by jointly minimizing the classification loss $\mathcal{L}_{\text{cls}}$, the reconstruction loss $\mathcal{L}_{\text{r}}$ computed based on real faces only, and the metric-learning loss $\mathcal{L}_{\text{m}}$.

## 3. Proposed Method

To capture the essential discrepancy between real and fake faces, we design a novel framework named RECCE, which consists of three main schemes, *i.e.*, reconstruction learning, multi-scale graph reasoning, and reconstruction guided attention, as illustrated in Figure 2. The reconstruction network aims to only model the distributions of real face images. As such, the learned representations are more likely to detect unknown forgery patterns. Moreover, to further mine the essentially discriminative representation, the multi-scale graph reasoning scheme aggregates the captured discrepancy information between real and fake faces in both the encoder and decoders of the reconstruction network in a multi-scale manner. Meanwhile, the reconstruction guided attention module guides the classification network to pay more attention to forgery traces. The following subsections present these three schemes in detail.

### 3.1. Reconstruction Learning

Since face forgery methods are always diverse, we argue that exploring the common characteristics of genuine faces is more suitable than overfitting specific forgery patterns presented in the training set. As such, we propose to perform reconstruction learning to restore real facial images only. To be specific, given an input image $\mathbf{X} \in \mathbb{R}^{h \times w \times 3}$, we train a reconstruction network $\mathcal{F}$ based on the encoder-decoder structure. As previous studies [57] have demonstrated that a plain reconstruction branch for restoring the original inputs would not significantly improve the learned representations, we add some white noises to the input samples during training to get $\tilde{\mathbf{X}}$, aiming to learn robust representations for real faces. Thus, the image reconstruction process can be formulated as:

$$\hat{\mathbf{X}} = \mathcal{F}(\tilde{\mathbf{X}}). \tag{1}$$

During the reconstruction process, we compute the reconstruction loss $\mathcal{L}_{\text{r}}$ between input real images and their reconstructed versions in a mini-batch as:

$$\mathcal{L}_{\text{r}} = \frac{1}{|R|} \sum_{i \in R} \|\hat{\mathbf{X}}_i - \mathbf{X}_i\|_1, \tag{2}$$

where $R$ denotes the set of real samples in a mini-batch and $|R|$ is the cardinality of $R$.

In addition to the reconstruction difference, we use a metric-learning loss to make real images close while real and fake images faraway in the embedding space. For simplicity, let $\mathbf{F} \in \mathbb{R}^{h' \times w' \times c}$ denote the output features of an encoder or decoder block. We apply the global average pooling operation to $\mathbf{F}$ and obtain the feature vector $\bar{\mathbf{F}} \in \mathbb{R}^c$ for each input sample. The metric-learning loss is:

$$\mathcal{L}_{\text{m}} = \frac{1}{N_{RR}} \sum_{i \in R, j \in R} d(\bar{\mathbf{F}}_i, \bar{\mathbf{F}}_j) - \frac{1}{N_{RF}} \sum_{i \in R, j \in F} d(\bar{\mathbf{F}}_i, \bar{\mathbf{F}}_j), \tag{3}$$

where $R, F$ denote the set of real and fake samples. $N_{RR}$ and $N_{RF}$ are the total number of (real, real) pairs and (real, fake) pairs, respectively. $d(\cdot, \cdot)$ is a pair-wise distant function based on the cosine distance:

$$d(\mathbf{a}, \mathbf{b}) = \frac{1 - \frac{\mathbf{a}}{\|\mathbf{a}\|_2} \cdot \frac{\mathbf{b}}{\|\mathbf{b}\|_2}}{2}. \tag{4}$$

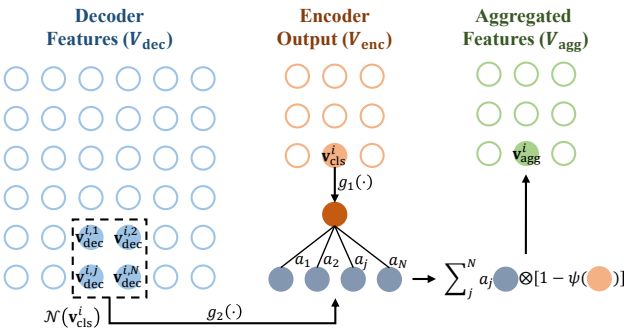


Figure 3. Illustration of the proposed multi-scale graph reasoning scheme to aggregate information in the encoder output (orange) and decoder features for a given scale (blue) to produce richer representations (green). This figure is best viewed in color.

The first part in $\mathcal{L}_{\text{m}}$ encourages learning compact representations from genuine faces while the second part emphasizes the differences between real and fake samples. Note that, different from the classic metric-learning loss [3, 21, 37, 48] which is directly applied to the feature extractor, our proposed loss is specially used to enhance the reconstruction difference to facilitate the reconstruction learning. Additionally, we do not constrain the compactness for fake data as their features vary substantially among different forgery techniques. We apply the metric-learning loss to the output of the last encoder block and each decoder block.

### 3.2. Multi-scale Graph Reasoning

When applying the metric-learning loss to the decoder, the useful information to separate real and fake images is embedded in the decoder as well. To effectively exploit the forgery clues captured by decoder features for final classification, we propose a multi-scale graph reasoning (MGR) module which combines the latent features of the decoder blocks and the encoder output into a bipartite graph for reasoning about forgery cues comprehensively.

Here, we take the feature maps of a decoder block for a given scale for description. As shown in Figure 3, we model the encoder output and the decoder features, *i.e.*, $\mathbf{F}_{\text{enc}}, \mathbf{F}_{\text{dec}}$, as two vertex set $V_{\text{enc}} = \{\mathbf{v}^i_{\text{enc}}\}_{i=1}^{h_1 \times w_1}$, $V_{\text{dec}} = \{\mathbf{v}^i_{\text{dec}}\}_{i=1}^{h_2 \times w_2}$, where each vertex represents a corresponding embedding vector of the original feature maps. $\mathcal{N}(\mathbf{v}^i_{\text{enc}}) = \{\mathbf{v}^{i,j}_{\text{dec}}\}_{j=1}^{N}$ denotes the set of vertices in $V_{\text{dec}}$ which is linked to $\mathbf{v}^i_{\text{enc}}$. $N$ is the number of vertices in the set. Concretely, graph reasoning aggregates the information from $\mathcal{N}(\mathbf{v}^i_{\text{enc}})$ to enrich the feature representations of $\mathbf{v}^i_{\text{enc}}$ for better reasoning about forgery cues. We keep the spatial correspondence when aggregating the information from the decoder to the encoder to model the local relationship since forgery traces usually reside in continuous local areas. As shown in Figure 3, the neighborhood of the orange solid vertex is the blue solid vertices in the dotted box. Given $\mathbf{v}^i_{\text{enc}}, \mathbf{v}^{i,j}_{\text{dec}}$, we first project them to a shared embedding space with two neural nets, $g_1(\cdot)$ and $g_2(\cdot)$, to get $\widetilde{\mathbf{v}}^i_{\text{enc}}, \widetilde{\mathbf{v}}^{i,j}_{\text{dec}}$, respectively. Next, we compute a weight coefficient $a_j$ to indicate the importance of $\mathbf{v}^{i,j}_{\text{dec}}$ to $\mathbf{v}^i_{\text{enc}}$. Particularly, we first concatenate the vertices from the two sub-graphs, and then passing through a single-layer network $\phi$ to get $a_j$ as:

$$a_j = \frac{\exp\left(\phi(\widetilde{\mathbf{v}}^i_{\text{enc}} \| \widetilde{\mathbf{v}}^{i,j}_{\text{dec}})\right)}{\sum_{\mathbf{v}^{i,l}_{\text{dec}} \in \mathcal{N}(\mathbf{v}^i_{\text{enc}})} \exp\left(\phi(\widetilde{\mathbf{v}}^i_{\text{enc}} \| \widetilde{\mathbf{v}}^{i,l}_{\text{dec}})\right)}, \tag{5}$$

where $\|$ denotes the concatenation operation. We then compute a $[0, 1]$-valued vector based on $\mathbf{v}^i_{\text{enc}}$ using a non-linear transformation $\psi(\cdot)$ to generate a feature richness measurement for $\widetilde{\mathbf{v}}^i_{\text{enc}}$ in the channel level. During information aggregation, we particularly enhance the channels of $\widetilde{\mathbf{v}}^{i,j}_{\text{dec}}$ when the weight of the corresponding channels of $\widetilde{\mathbf{v}}^i_{\text{enc}}$ is small. The aggregated feature vector $\mathbf{v}^i_{\text{agg}}$ is computed by:

$$\mathbf{v}^i_{\text{agg}} = \sum_{j=1}^{N} a_j \widetilde{\mathbf{v}}^{i,j}_{\text{dec}} \otimes \left[1 - \psi(\mathbf{v}^i_{\text{enc}})\right], \tag{6}$$

where $\otimes$ is the element-wise multiplication.

Since different manipulation techniques result in forged traces across scales, we propose to mine the forgery information in a multi-scale manner to obtain comprehensive representations. To be specific, the output features of the encoder aggregate each block output of the decoder in a multi-scale manner. The aggregated features $\{\mathbf{v}^i_{\text{agg}}\}$ in different scales are concatenated with $\mathbf{v}^i_{\text{enc}}$ and then pass through a sigmoid function followed by two fully-connected layers to produce the enhanced feature vector $\mathbf{v}^i_{\text{enh}}$ with the same channel dimension as $\mathbf{v}^i_{\text{enc}}$. Finally, $\mathbf{v}^i_{\text{enh}}$ are assembled spatially to obtain the enhanced feature maps $\mathbf{F}_{\text{enh}}$ for the following reconstruction guided attention.

### 3.3. Reconstruction Guided Attention

Equipped with the constraints of the reconstruction network, the reconstructed forged faces largely differ from the input forged faces in visual appearance. This motivates us to use the reconstruction difference to indicate the probably manipulated traces. To this end, we propose the reconstruction guided attention (RGA) module, which pays more attention to the probable forgery regions to facilitate later classification.

As shown in Figure 2, given the reconstructed image $\widehat{\mathbf{X}}$ and the original image $\mathbf{X}$, we first compute their difference in pixel level to get the difference mask $\mathbf{m}$ as

$$\mathbf{m} = \left|\widehat{\mathbf{X}} - \mathbf{X}\right|, \tag{7}$$

where $|\cdot|$ refers to the absolute value function. Given $\mathbf{F}_{\text{enh}}$ the enhanced feature maps mentioned in Section 3.2, we

| Methods | FF++ (c23) | | FF++ (c40) | | Celeb-DF | | WildDeepfake | |
|---|---|---|---|---|---|---|---|---|
| | Acc (%) | AUC (%) | Acc (%) | AUC (%) | Acc (%) | AUC (%) | Acc(%) | AUC (%) |
| MesoNet [1] | 83.10 | – | 70.47 | – | – | – | 64.47 | – |
| Multi-task [30] | 85.65 | 85.43 | 81.30 | 75.59 | – | – | – | – |
| Xception [35] | 95.73 | 96.30 | 86.86 | 89.30 | 97.90 | 99.73 | 77.25 | 86.76 |
| Face X-ray [23] | – | 87.40 | – | 61.60 | – | – | – | – |
| Two-branch [29] | 96.43 | 98.70 | 86.34 | 86.59 | – | – | – | – |
| SPSL [25] | 91.50 | 95.32 | 81.57 | 82.82 | – | – | – | – |
| RFM [45] | 95.69 | 98.79 | 87.06 | 89.83 | 97.96 | **99.94** | 77.38 | 83.92 |
| Freq-SCL [22] | 96.69 | 99.28 | 89.00 | 92.39 | – | – | – | – |
| Add-Net [60] | 96.78 | 97.74 | 87.50 | 91.01 | 96.93 | 99.55 | 76.25 | 86.17 |
| $F^3$-Net [33] | 97.52 | 98.10 | 90.43 | 93.30 | 95.95 | 98.93 | 80.66 | 87.53 |
| MultiAtt [55] | **97.60** | 99.29 | 88.69 | 90.40 | 97.92 | **99.94** | 82.86 | 90.71 |
| RECCE (Ours) | 97.06 | **99.32** | **91.03** | **95.02** | **98.59** | **99.94** | **83.25** | **92.02** |

Table 1. Intra-testing comparisons. The proposed method performs favorably over current state-of-the-art approaches.

compute the attention map based on the difference mask and apply it to $\mathbf{F}_{\text{enh}}$ spatially to get $\mathbf{F}'_{\text{enh}}$. Then, we add $\mathbf{F}'_{\text{enh}}$ and $\mathbf{F}_{\text{enh}}$ to obtain the attended output features:

$$\mathbf{F}'_{\text{enh}} = \sigma\left(f_1(\mathbf{m})\right) \otimes f_2(\mathbf{F}_{\text{enh}}), \tag{8}$$

$$\mathbf{F}_{\text{att}} = \mathbf{F}'_{\text{enh}} + \mathbf{F}_{\text{enh}}, \tag{9}$$

where $f_1, f_2$ represent the convolutional operations, $\sigma$ is the sigmoid function, and $\otimes$ denotes the element-wise multiplication. For simplicity, we omit the spatial size of these tensors and use the bilinear interpolation to keep the spatial size properly for mentioned operations.

### 3.4. Loss Function

The total loss function $\mathcal{L}$ of the proposed framework includes the reconstruction loss and the metric-learning loss for reconstruction learning, together with the cross-entropy loss $\mathcal{L}_{\text{cls}}$ for binary classification:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{r}} + \lambda_2 \mathcal{L}_{\text{m}}, \tag{10}$$

where $\lambda_1$ and $\lambda_2$ are weight parameters for balancing different losses.

## 4. Experiments

### 4.1. Experimental Setup

**Datasets.** We evaluate our proposed method and existing approaches on FaceForensics++ (FF++) [35], Celeb-DF [24], WildDeepfake (WDF) [60] and DFDC [9]. FF++ is the most widely used dataset containing four types of manipulation techniques, *i.e.*, Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Celeb-DF includes 590 real videos and 5,639 high-quality fake videos which are crafted by the improved DeepFake algorithm [24]. WildDeepfake is a real-world dataset that contains 3,805 real sequences and 3,509 fake sequences. All the videos in it are obtained from the internet with more identities presented in various scenes. DFDC is a large-scale dataset which contains 128,154 facial videos of 960 subjects. Due to the variety of manipulations and perturbations, it is very challenging for the existing methods.

**Evaluation Metrics.** To evaluate our method, we report the most commonly used metrics in related arts [1, 6, 22, 33, 35, 55, 59], including Accuracy (Acc), Area Under the Receiver Operating Characteristic Curve (AUC), and Equal Error Rate (EER). We also report *LogLoss* on DFDC, consistent with the setting of its corresponding contest [9].

**Implementation Details.** The proposed framework is implemented based on the Xception [7]. We train it with a batch size of 32, the Adam [19] optimizer with an initial learning rate of 2e-4 and a weight decay of 1e-5. A step learning rate scheduler is used to adjust the learning rate. $\lambda_1$ and $\lambda_2$ in Equation (10) are empirically set to 0.1. We only use random horizontal flipping for data augmentation.

### 4.2. Experimental Results

**Intra-testing.** In this section, we compare our proposed method with current state-of-the-art approaches. As shown in Table 1, for FF++ dataset, our method consistently achieves great improvements under different quality settings. Especially on the challenging c40 (low-quality) setting, compared with $F^3$-Net [33], the AUC score of our method exceeds it by 1.72%. To explain, over-compression destroys the frequency clues that $F^3$-Net relies upon, while our approach yields a more robust representation through reconstruction learning that serves as effective guidance for forgery classification. Note that, though MultiAtt [55] equipped with EfficientNet-b4 reaches the highest Acc on FF++ c23 (high-quality), our method based on Xception still achieves comparable results and exceeds it by a large margin on the low-quality setting. Different from Multi-task [30] which employs reconstruction constraints for both real and fake faces, the proposed RECCE framework only

| Methods | Acc (%) ↑ | AUC (%) ↑ | LogLoss ↓ |
|---|---|---|---|
| Xception [35] | 79.35 | 89.50 | 0.4916 |
| RFM [45] | 80.83 | 89.75 | 0.5810 |
| Add-Net [60] | 78.71 | 89.85 | 0.5072 |
| $F^3$-Net [33] | 76.17 | 88.39 | 0.5196 |
| MultiAtt [55] | 76.81 | 90.32 | 0.5291 |
| RECCE (Ours) | **81.20** | **91.33** | **0.4341** |

Table 2. Intra-testing comparisons on the DFDC [9] dataset.

models the distribution of real samples and promotes comprehensive difference information. Thus, our method significantly outperforms the counterpart. The performance gains can also be observed on Celeb-DF and the realistic dataset WildDeepfake, while in the latter our method reaches a state-of-the-art result by improving the Acc by 0.39% and the AUC by 1.31%. The above results demonstrate the effectiveness of the proposed RECCE framework.

Furthermore, we evaluate our method on the challenging DFDC dataset. Since few existing arts report the performance on it, we re-implement several state-of-the-art methods for a fair comparison, including RFM [45], Add-Net [60], $F^3$-Net [33] and MultiAtt [55]. As shown in Table 2, our method outperforms other approaches by 0.37% and 1.01% in terms of Acc and AUC, while LogLoss decreases by 0.0575. These results validate the effectiveness of our proposed method under complicated scenarios.

**Cross-testing.** To evaluate the generalization ability of our method on unknown forgeries, we conduct cross-dataset experiments by training and testing on different datasets. Specifically, we train the models on FF++ c40, and then test them on WildeDeepfake, Celeb-DF, and DFDC, respectively. The results are shown in Table 3. From the table, we observe that RECCE generally outperforms all listed methods on unseen test data, often by a large margin. For instance, when testing on WildDeepfake dataset, the AUC score of most previous methods drop to around 60%. Differently, RECCE reaches an AUC of 64.31%, which exceeds MultiAtt [55] by 4.57%. The performance mainly benefits from the proposed RECCE framework which only models the distribution of real faces, while MGR and RGA guide the model to learn essential discrepancy between real and fake faces. Instead of overfitting with specific forged patterns as in existing methods, our method treats all unknown forgery types as outliers to achieve better generalizability.

We further conduct a fine-grained cross-testing by training on a specific manipulation technique and testing on the others listed in FF++ c40. We compare our method with approaches that focus on specific forgery patterns, *e.g.*, Freq-SCL [22] and MultiAtt [55], in Table 4. Our method generally outperforms others on unseen forgery types. These results verify that it is feasible to explore common features of real faces to distinguish forgeries with unknown patterns.

| Methods | WDF | | Celeb-DF | | DFDC | |
|---|---|---|---|---|---|---|
| | AUC ↑ | EER ↓ | AUC ↑ | EER ↓ | AUC ↑ | EER ↓ |
| Xception [35] | 62.72 | 40.65 | 61.80 | 41.73 | 63.61 | 40.58 |
| RFM [45] | 57.75 | 45.45 | 65.63 | 38.54 | 66.01 | 39.05 |
| Add-Net [60] | 62.35 | 41.42 | 65.29 | 38.90 | 64.78 | 40.23 |
| $F^3$-Net [33] | 57.10 | 45.12 | 61.51 | 42.03 | 64.60 | 39.84 |
| MultiAtt [55] | 59.74 | 43.73 | 67.02 | 37.90 | 68.01 | 37.17 |
| RECCE (Ours) | **64.31** | **40.53** | **68.71** | **35.73** | **69.06** | **36.08** |

Table 3. Cross-testing in terms of AUC (%) and EER (%) by training on FF++ [35].

| Methods | Train | DF | F2F | FS | NT | Cross Avg. |
|---|---|---|---|---|---|---|
| Freq-SCL [22] | DF | 98.91 | 58.90 | 66.87 | 63.61 | 63.13 |
| MultiAtt [55] | | 99.51 | 66.41 | 67.33 | 66.01 | 66.58 |
| RECCE (Ours) | | **99.65** | **70.66** | **74.29** | **67.34** | **70.76** |
| Freq-SCL [22] | F2F | 67.55 | 93.06 | 55.35 | 66.66 | 63.19 |
| MultiAtt [55] | | 73.04 | 97.96 | **65.10** | 71.88 | 70.01 |
| RECCE (Ours) | | **75.99** | **98.06** | 64.53 | **72.32** | **70.95** |
| Freq-SCL [22] | FS | 75.90 | 54.64 | 98.37 | 49.72 | 60.09 |
| MultiAtt [55] | | 82.33 | 61.65 | **98.82** | 54.79 | 66.26 |
| RECCE (Ours) | | **82.39** | **64.44** | **98.82** | **56.70** | **67.84** |
| Freq-SCL [22] | NT | **79.09** | 74.21 | 53.99 | 88.54 | 69.10 |
| MultiAtt [55] | | 74.56 | 80.61 | 60.90 | 93.34 | 72.02 |
| RECCE (Ours) | | 78.83 | **80.89** | **63.70** | **93.63** | **74.47** |

Table 4. Cross-testing in terms of AUC (%) on different manipulation techniques. Gray background means within-dataset results.

**Reconstruction visualization.** For an intuitive understanding of reconstruction learning, we visualize the outputs of the reconstruction network and the difference between original input, as shown in Figure 4. We can see that the real faces can be well reconstructed with little blur, while the forged regions of fake ones cannot be restored. The difference masks further display the discrepancy between real and forged faces, indicating possible traces of forged region, even if our method is only trained under image-level supervision. Taking the NeuralTextures (NT) method as an example, which operates only on the mouth region, the difference masks of the corresponding samples show larger values exactly around the mouth area. Moreover, for the realistic WildDeepfake dataset, though the source and manipulation method remain unknown, our method can still indicate possibly forged regions. The visualization validates that the proposed framework can effectively capture the essential discrepancy between real and fake faces.

### 4.3. Ablation Study

**Effectiveness of proposed components.** In this part, we conduct the ablation study on different components proposed in our framework to evaluate their effectiveness. Specifically, we develop the following variants: (a)

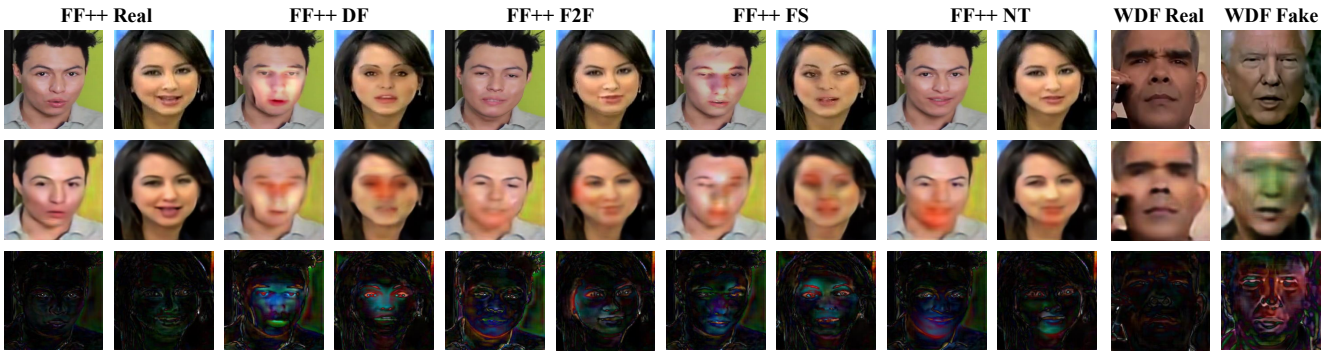


Figure 4. Reconstruction visualization of the proposed method on the FaceForensics++ [35] dataset and WildDeepfake [60] dataset. The first row displays the input images. The second row and the third row show reconstruction results and pixel-level differences, respectively.

| ID | Rec. | MGR | RGA | Acc (%) | AUC (%) |
|---|---|---|---|---|---|
| (a) | | | | 77.25 | 86.76 |
| (b) | √ | | | 81.19 | 89.61 |
| (c) | √ | √ | | 81.48 | 91.10 |
| (d) | √ | | √ | 82.15 | 89.71 |
| RECCE | √ | √ | √ | **83.25** | **92.02** |

Table 5. Effectiveness of the proposed components in our method on the WildDeepfake [60] dataset.

| ID | $\mathcal{L}_{\mathrm{r}}$ | $\mathcal{L}_{\mathrm{m}}$ | Acc (%) | AUC (%) |
|---|---|---|---|---|
| (a) | real & fake | √ | 80.62 | 88.92 |
| (b) | real | – | 81.36 | 90.49 |
| RECCE | real | √ | **83.25** | **92.02** |

Table 6. Effectiveness of the proposed constraints in our method on the WildDeepfake [60] dataset.

the baseline model which follows the classic image classification pipeline, *i.e.*, Xception [35], (b) the baseline model equipped with the proposed reconstruction learning scheme, (c) the proposed method without MGR, and (d) the proposed method without RGA. The quantitative results are listed in Table 5. Comparing variants (a) and (b), we can see that the proposed reconstruction learning brings 3.94% Acc and 2.85% AUC gains over the baseline method. Solely employing the variant (b) with the MGR module that enhances the learning of classification side with the comprehensive representations captured by the decoder, the resulting variant (c) obtains a performance gain by a large margin, *i.e.* 1.49% on AUC. From variants (b) and (d), we observe an improvement on both Acc and AUC metrics when adding the RGA module which highlights the probably forged regions on the output of the encoder. The best performance is achieved when combining all the proposed components with Acc and AUC of 83.25% and 92.02% respectively.

**Effectiveness of proposed constraints.** To investigate the effectiveness of the proposed constraints used in the reconstruction network, we design two variants of our method: (a) reconstruction loss $\mathcal{L}_{\mathrm{r}}$ in Equation (2) is computed for both real and fake faces, and (b) our method without the metric-learning loss $\mathcal{L}_{\mathrm{m}}$ in Equation (3). The results are presented in Table 6. Comparing variant (a) and our method, we observe that training the reconstruction network on both real and fake images hampers the model from learning a unified representation for real ones. Regarding variant (b) and our method, we find that $\mathcal{L}_{\mathrm{m}}$ brings a 1.53% AUC gain. This is mainly because $\mathcal{L}_{\mathrm{m}}$ makes real images closer and pushes away real and fake ones in the embedding space. These results demonstrate that the proposed constraints are conducive to the discrepancy mining process.

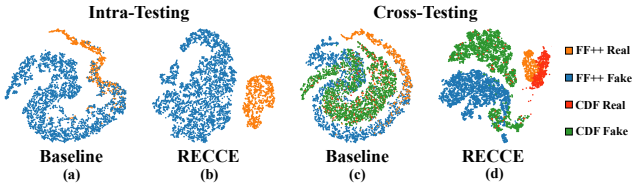


Figure 5. The t-sne [43] embedding visualization of the features encoded in the baseline method and RECCE. Best viewed in color.

## 4.4. Experimental Analysis

**Analysis of feature distribution.** In this section, we visualize the learned feature distribution of the baseline [35] and our approach trained on FF++ c40 using t-sne [43]. The features of our method are extracted from the layer right before the last fully-connected layer, and the results are shown in Figure 5. In particular, we randomly sample 5000 images from FF++ for the intra-testing setting (*i.e.*, (a) and (b)), and additionally select 5000 samples from Celeb-DF for the cross-testing setting (*i.e.*, (c) and (d)). From the figure, we observe that the baseline is short of the understanding of forgeries as the clusters of real and fake images are indistinguishable. In contrast, our method embeds the real faces into a relatively compact feature space, whether on known

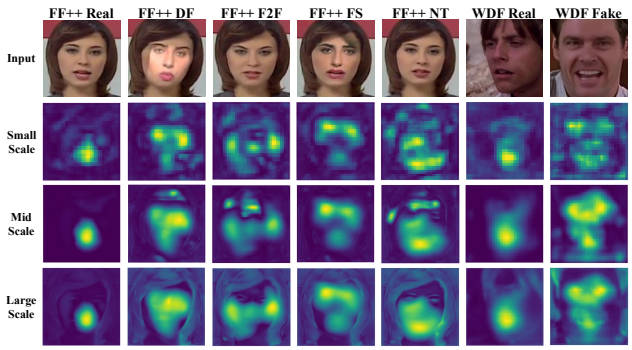


Figure 6. The visualization of the feature maps in the decoder of the reconstruction net at different output scales.

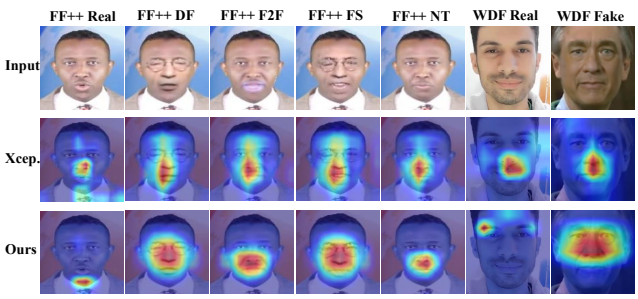


Figure 7. The Grad-CAM [38] visualization.

| Methods | Compress | Blur | Contrast | Saturate | Pixelate | Avg. |
|---|---|---|---|---|---|---|
| Xception [35] | 86.01 | 78.29 | 81.90 | 84.96 | 66.24 | 79.48 |
| RFM [45] | 83.74 | 75.34 | 79.77 | 82.59 | 71.25 | 78.54 |
| Add-Net [60] | 83.34 | 79.66 | 84.46 | 85.13 | 64.33 | 79.38 |
| $F^3$-Net [33] | 86.71 | 78.99 | 86.53 | 87.67 | 73.23 | 82.63 |
| MultiAtt [55] | 89.64 | 80.98 | 89.30 | 90.37 | 79.44 | 85.95 |
| RECCE (Ours) | **89.65** | **87.29** | **91.19** | **91.74** | **83.88** | **88.75** |

Table 7. Robustness evaluation in terms of AUC (%) on WildDeepfake [60] dataset. "Avg." indicates the mean score.

or unknown data, which suggests that our model captures the common representations of genuine faces. Besides, the clusters for real samples and fake samples are separated by an obvious margin, which reveals the discrepancy between genuine and forged faces. The visual results, from another viewpoint, validate the effectiveness of our approach that focuses on genuine faces to capture the differences in faces.

**Analysis of multi-scale features in the decoder.** In this section, we visualize the feature maps from different layers of the decoder. The results are shown in Figure 6. From the figure, we observe that the decoder features at different scales are activated with different intensities. On one hand, the forgery clues at large-scale feature maps are more comprehensive and richer, but they also contain some irrelevant background noise. On the other hand, the forgery clues at small-scale feature maps are fine-grained but incomplete. Therefore, combining the multi-scale information is beneficial for the model to attend to the substantial difference while avoiding the interference of irrelevant noise.

**Analysis of classification decision.** To better understand the decision-making mechanism of our method, we provide the Grad-CAM [38] visualization on FF++ in Figure 7. We observe that the baseline method mainly focuses on the central region of images for classification regardless of the facial authenticity, lacking a comprehensive understanding of different forgeries. Differently, our method generates distinguishable heatmaps for real and fake faces, where the prominent regions vary in forgery techniques, even though it only uses binary labels for training. For instance, both of the heatmaps for DeepFakes (DF) and FaceSwap (FS) focus on the main facial area while that for NeuralTextures (NT) localizes mouth regions. The results explain the effectiveness of RECCE from the decision-making perspective.

**Analysis of robustness.** Considering the ubiquity of image processing on social media, we investigate the performance under several perturbations suggested by [15, 17], *i.e.*, image compression, Gaussian blur, contrast jitter, saturate jitter, and pixelation. We show the results in Table 7. We can see that RECCE is more robust to the listed perturbations than exiting approaches. It is worth noticing that previous methods undergo an obvious performance drop when encountering Gaussian blur (which destroys the frequency statistics) and pixelation (which drops the texture information). The degradation indicates that emphasizing specific forgery patterns presented in training data is vulnerable to common perturbations. However, our method outperforms them by a large margin, *i.e.*, 6.31% for Gaussian blur and 4.44% for pixelation. In average, our model obtains 2.80% AUC gain over the state-of-the-art MultiAtt [55], which demonstrates the robustness of our proposed method.

## 5. Conclusion

In this paper, we propose a new perspective for face forgery detection that focuses on common compact representations of real faces to learn the discrepancy between real and forged faces. The innovative multi-scale graph reasoning module combines encoder output and decoder features into bipartite graphs in a multi-scale fashion for reasoning about forgery clues. Meanwhile, the reconstruction guided attention module is introduced to guide the model to focus on possibly forgery traces. Extensive experiments and detailed visualizations validate the robustness and generalizability of our method on widely-used benchmark datasets.

**Acknowledgements.** This work was supported by NSFC (61906119, U19B2035), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and CCF-Tencent Open Research Fund.

## References


[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: A compact facial video forgery detection network. In *WIFS*, 2018.

[2] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter N. Belhumeur, and Shree K. Nayar. Face swapping: Automatically replacing faces in photographs. *ACM Trans. Graph.*, 27(3):39, 2008.

[3] Shenhao Cao, Qin Zou, Xiuqing Mao, Dengpan Ye, and Zhongyuan Wang. Metric learning for anti-compression facial forgery detection. In *ACM MM*, 2021.

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. In *ICCV*, 2019.

[5] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-supervised learning. *IEEE Trans. Neural Networks*, 20(3):542, 2009.

[6] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *AAAI*, 2021.

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

[8] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. On the detection of digital face manipulation. In *CVPR*, 2020.

[9] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[10] Mengnan Du, Shiva K. Pentyala, Yuening Li, and Xia Hu. Towards generalizable deepfake detection with locality-aware autoencoder. In *CIKM*, 2020.

[11] Yue Gao, Fangyun Wei, Jianmin Bao, Shuyang Gu, Dong Chen, Fang Wen, and Zhouhui Lian. High-fidelity and arbitrary face editing. In *CVPR*, 2021.

[12] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *AAAI*, 2022.

[13] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *ACM MM*, 2021.

[14] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *AAAI*, 2022.

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

[16] Zhizhong Han, Xiyang Wang, Yu-Shen Liu, and Matthias Zwicker. Multi-angle point cloud-vae: Unsupervised feature learning for 3d point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. In *ICCV*, 2019.

[17] 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*, 2020.

[18] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *ICCV*, 2021.

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

[20] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *ICCV*, 2017.

[21] Akash Kumar, Arnav Bhavsar, and Rajesh Verma. Detecting deepfakes with metric learning. In *International Workshop on Biometrics and Forensics*, 2020.

[22] 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*, 2021.

[23] 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*, 2020.

[24] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *CVPR*, 2020.

[25] 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*, 2021.

[26] Xinhai Liu, Xinchen Liu, Zhizhong Han, and Yu-Shen Liu. Spu-net: Self-supervised point cloud upsampling by coarse-to-fine reconstruction with self-projection optimization. *arXiv preprint arXiv:2012.04439*, 2020.

[27] Siwei Lyu. Deepfake detection: Current challenges and next steps. In *ICME Workshop*, 2020.

[28] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. In *ICML*, 2016.

[29] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, 2020.

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

[31] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP*, 2019.

[32] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[33] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020.

[34] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NeurIPS*, 2015.

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

[36] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *ICLR*, 2020.

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *CVPR*, 2017.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[40] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *AAAI*, 2022.

[41] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13, 2017.

[42] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4):66:1–66:12, 2019.

[43] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[44] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

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

[46] Xinyao Wang, Taiping Yao, Shouhong Ding, and Lizhuang Ma. Face manipulation detection via auxiliary supervision. In *ICONIP*, 2020.

[47] Zhuhui Wang, Shijie Wang, Haojie Li, Zhi Dou, and Jianjun Li. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In *AAAI*, 2020.

[48] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[49] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *CVPR*, 2021.

[50] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. ReenactGAN: Learning to reenact faces via boundary transfer. In *ECCV*, 2018.

[51] Burhaneddin Yaman, Chetan Shenoy, Zilin Deng, Steen Moeller, Hossam El-Rewaidy, Reza Nezafat, and Mehmet Akçakaya. Self-supervised physics-guided deep learning reconstruction for high-resolution 3D LGE CMR. In *International Symposium on Biomedical Imaging*, 2021.

[52] Jie Yang, Yong Shi, and Zhiquan Qi. DFR: Deep feature reconstruction for unsupervised anomaly segmentation. *arXiv preprint arXiv:2012.07122*, 2020.

[53] Guangming Yao, Yi Yuan, Tianjia Shao, Shuang Li, Shanqi Liu, Yong Liu, Mengmeng Wang, and Kun Zhou. One-shot face reenactment using appearance adaptive normalization. In *AAAI*, 2021.

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

[55] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, 2021.

[56] Yifan Zhao, Ke Yan, Feiyue Huang, and Jia Li. Graph-based high-order relation discovery for fine-grained recognition. In *CVPR*, 2021.

[57] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *ICCV*, 2021.

[58] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection. In *CVPR Workshops*, 2017.

[59] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z. Li. Face forgery detection by 3D decomposition. In *CVPR*, 2021.

[60] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. WildDeepfake: A challenging real-world dataset for deepfake detection. In *ACM MM*, 2020.