# Controllable Guide-Space for Generalizable Face Forgery Detection

Ying Guo *, Cheng Zhen *, Pengfei Yan †
Vision AI Department, Meituan
{guoying16, zhencheng02, yanpengfei03}@meituan.com

## Abstract

*Recent studies on face forgery detection have shown satisfactory performance for methods involved in training datasets, but are not ideal enough for unknown domains. This motivates many works to improve the generalization, but forgery-irrelevant information, such as image background and identity, still exists in different domain features and causes unexpected clustering, limiting the generalization. In this paper, we propose a controllable guide-space (GS) method to enhance the discrimination of different forgery domains, so as to increase the forgery relevance of features and thereby improve the generalization. The well-designed guide-space can simultaneously achieve both the proper separation of forgery domains and the large distance between real-forgery domains in an explicit and controllable manner. Moreover, for better discrimination, we use a decoupling module to weaken the interference of forgery-irrelevant correlations between domains. Furthermore, we make adjustments to the decision boundary manifold according to the clustering degree of the same domain features within the neighborhood. Extensive experiments in multiple in-domain and cross-domain settings confirm that our method can achieve state-of-the-art generalization.*

## 1. Introduction

Face forgery technology [1, 2, 13] has made vigorous development in recent years. However, these realistic forgery faces are sometimes abused to maliciously disguise identities, especially celebrities and politicians, causing serious social problems. Therefore, how to reduce this risk has attracted widespread attention from researchers.

Convolutional neural networks (CNNs) have shown excellent performance in face forgery detection [10, 49, 32, 12]. According to forgery or not, this task is often formalized as a binary classification problem, and some suitable classification networks [32, 30] are introduced to this task. Although they perform well in the training domain, the learned features may be method-specific for the forgery methods within the training set [5], and cannot show satisfactory generalization in unknown forgery methods.

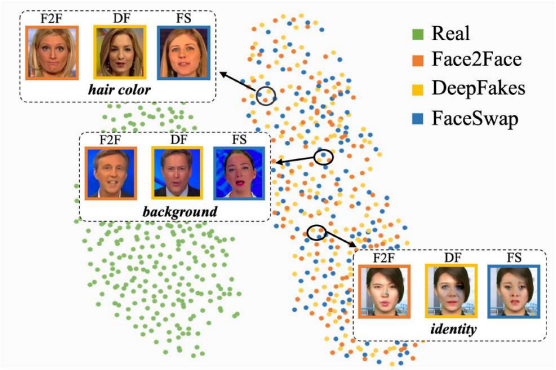


Figure 1. Limitations of existing methods: features from different domains are still clustered by forgery-irrelevant similarities (image background, hair color, identity, etc.), proving that features still contain forgery-irrelevant information, limiting the generalization.

Forgery data generated from various methods corresponds to different forgery domains. To improve the generalization in unknown domains, some works [22, 34, 6] study the common artifacts of various forgeries and use data augmentation to synthesize more training data. Several other works are devoted to mining better discriminative features, and attention mechanisms [47], local relation [7], and frequency information [21, 31] are also introduced to capture better forgery traces. SRM [26] suppresses the acquisition of color and texture via high-frequency noise, thereby solving the overfitting to the training data. In addition, RECCE [5] copes with the complexity of various forgery domains by learning compact real representations based on the reconstruction. Compared with previous methods, the exploration of common characteristics and discriminative information allows the model to learn relatively more generalized features.

Nevertheless, during training, the model tries to increase the discrimination between real and fake features, but treats different forgery domains (i.e. different forgery methods) as the unified "fake" category without distinction. In common training sets [33, 23], there are some similarities between massive data in forgery-irrelevant information, such as hair color, image background, and identity. Due to the uniform

*Equal Contribution. †Corresponding author.

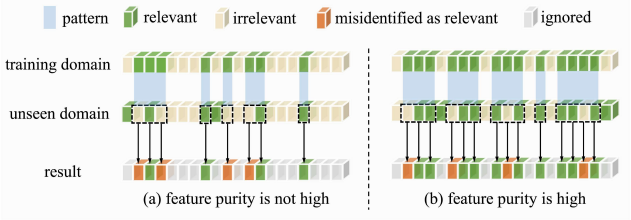


Figure 2. Comparisons of the forgery-relevance for model focusing results at different feature purities.

fake categorization, the goal of training is only to distinguish the fake from the real, without making further distinctions between forgery types. As a result, as shown in Figure 1, some features will present a clustering phenomenon based on above forgery-irrelevant similarities (hair color, background, identity, etc.) rather than forgery domain characteristics that are more relevant to the forgery detection task. This demonstrates that the learned features inevitably still contain some forgery-irrelevant information [16].

The mixing of irrelevant information in features (i.e. the feature purity is not high) may limit the generalization. As shown in Figure 2 (a), in the training domain, guided by the supervised information, the model will learn a pattern of which features are more relevant, and show good performance in the current domain. However, in unseen domains, the feature distribution has a deviation. Following the original pattern, the features that the model focuses on will contain a high proportion of irrelevant information, which may cause the model to make decisions based on the similarity of such irrelevant information. On the contrary, as shown in Figure 2 (b), when the feature purity is high, the irrelevant information contained in the feature itself is less. Even if the distribution of the unseen domain is biased, the proportion of irrelevant information in the extracted features will be correspondingly less. So we believe that higher feature purity will help the generalization, and we also give a theoretical proof in Appendix 1.

Based on above considerations, in this paper, instead of treating all forgery types as a unified category, we propose a novel guide-space (GS) based framework to increase a proper level of discrimination between different forgery domains. In this way, by learning the differences between forgery domains and the consistency of the same domain, the model can further pay more attention to forgery traces. And separating the features of different domains can reduce their correlation in irrelevant information. The learned features are more forgery-relevant, thus helping the generalization.

Specifically, the increase in forgery domain discrimination needs to be controlled within a certain range, because a larger real-forgery distance should be preferably maintained at the same time. In this way, in unseen domains, the forgery features will be located far away from real features with a higher probability. Thus in our guide-space, we construct the guide embeddings of the real and different forgery domains, and make the features approach their respective guide embeddings to actively control the compactness of the real domain and the separation degree between different forgery domains. Further, considering that the correlation between different domains in terms of forgery-irrelevant similarity will interfere with the domain distinction, we mine this potential correlation based on the clustering results of the self-supervised features of images, and decouple the irrelevant information accordingly. In addition, we design a decision boundary manifold adjustment module (A-DBM) based on the degree of feature aggregation, to better realize the feature distribution defined by the guide-space.

In summary, this paper has the following contributions:

- We argue that a proper level of discrimination between different forgery domains is also important to improve the generalization, so as to capture more forgery-relevant information and to weaken the impact of forgery-irrelevant information.
- We construct a guide-space to achieve the controllable separation of both real-forgery domains and forgery-forgery domains, and further decouple the forgery-irrelevant correlation between different domains to reduce their interference on domain separation.
- We design an adjustment strategy for the decision boundary manifold to make the features of the same domain better clustered and compliant with the distribution of the guide-space.
- Extensive experiments in multiple cross-domain settings confirm that our method can realize the state-of-the-art generalization, and achieve the cross-domain AUC of 84.97% and 81.65% on CelebDF and DFDC.

## 2. Related works

Face forgery detection based on convolutional neural networks (CNNs) has been widely used [49, 3, 14, 39, 27, 45, 19, 29, 50]. Early works [32, 30] apply suitable classification networks to forgery detection tasks, and achieve good performance on the forgery domains presented in the training set. However, the learned features may be more suitable for forgery methods presented in the training set [5], and cannot show good generalization on unknown forgery domains.

Recently, more efforts have been made to improve this generalization. Some methods attempt to learn common characteristics of different forgery domains. For example, works in [22, 34, 28, 6] use data augmentation to simulate common artifacts (blending boundary, color inconsistency, etc.) of forgeries. SPSL [24] captures the phase spectrum changes caused by common up-sampling operations during the forgery process. They exhibit improved generalization, but these common features often cover a limited variety of forgeries. Several other studies are devoted to mining better

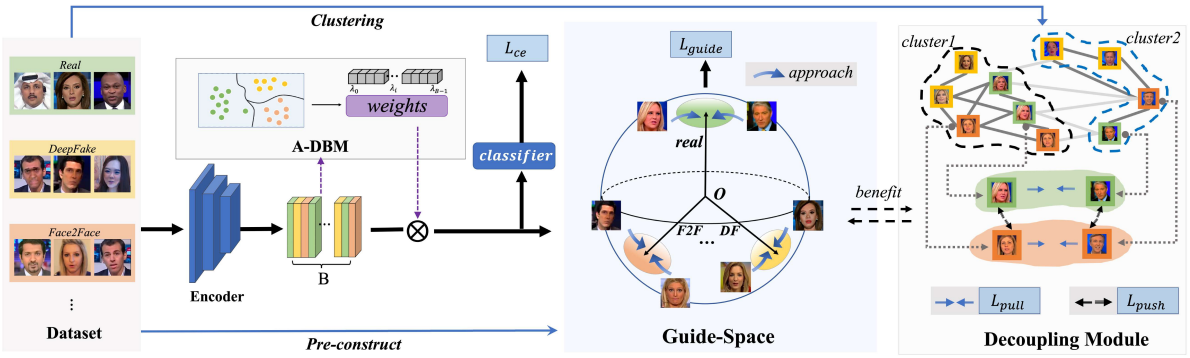


Figure 3. Overview of our framework, including guide-space based controllable optimization, adjustment of decision boundary manifold (A-DBM), and irrelevant information decoupling module.

discriminative features. To better capture forgery clues, the attention mechanism [47, 10, 44], amplification strategy [12], or local relationships [48, 7] between regions are studied. Besides low-level RGB features, the frequency information [31, 26, 17, 21] is also introduced. In addition, RECCE [5] based on reconstruction learning tries to learn compact representations of real data to cope with the complexity of forgery domains. LTW [37] utilizes meta-learning to balance the performance across multiple domains.

Ideally, for a generalized model that extracts forgery-related information, the forgery features should be aggregated according to their respective domain types rather than forgery-irrelevant information. However, although the above methods mine better forgery traces, they treat different forgery domains as the uniform "fake", which makes the features with forgery-irrelevant similarities but belonging to different domains still cluster together.

## 3. Methodology

To improve the generalization, we propose a novel guide-space (GS) based framework, which consists of three main schemes, i.e., guide-space based controllable optimization, adjustment of decision boundary manifold (A-DBM), and decoupling module for irrelevant information, as illustrated in Figure 3. We first pre-construct an ideal guide-space, making features closer to their guide embeddings of respective domains. To better aggregate features of the same domain, we adjust the decision boundary manifold by setting weights of samples within a batch. Further, to mitigate the interference of irrelevant correlations, we decouple these correlations with the aid of self-supervised feature clustering. The guide-space and decoupling module can benefit from each other to make features achieve better forgery relevance. The following subsections show details of three schemes.

### 3.1. Construction of guide-space

Before training, we first construct a guide-space containing guide embeddings for the real domain and different types of forgery domains. In the subsequent training, the features

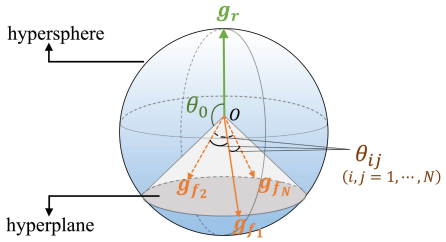


Figure 4. Visualization of the guide-space. The real guide embedding $\boldsymbol{g_r}$ and all forgery guide embeddings $\boldsymbol{g_f}$ are at a fixed angle $\theta_0$ ($\theta_0 > \theta_{ij}$), and the larger the angle $\theta_{ij}$ between $\boldsymbol{g_f}$, the better.

of different domains are approached to their respective guide embeddings to achieve distinguishability from each other.

The construction of the guide-space requires the dimension $d$ of the face feature representation and the number of forgery categories $N$ in the training set. The features lie on a hypersphere of unit length $S = \{v \in R^d \mid \|v\| = 1\}$. Let $\boldsymbol{g_r}$ and $\boldsymbol{g_f} = \left\{\boldsymbol{g}_{f_i} \mid i = 1,\cdots,N\right\}$ represent the guide embeddings of the real and forgery domains to be solved, respectively. The visualization of $\boldsymbol{g_r}$ and $\boldsymbol{g_f}$ in the guide-space is shown in Figure 4.

On the one hand, we make $\boldsymbol{g_r}$ and all embeddings in $\boldsymbol{g_f}$ present a fixed large angle $\theta_0$ (a hyperparameter) to separate real and fake features. $\theta_0$ can explicitly and actively control the separation degree between real-forgery domains, and a large $\theta_0$ ensures the compactness of real domain features:

$$e^{\boldsymbol{g_r}^T \boldsymbol{g}_{f_i}} = e^{cos(\theta_0)} (i = 1, \cdots, N) \tag{1}$$

On the other hand, the restriction of $\theta_0$ makes the embeddings in $\boldsymbol{g_f}$ be located in a hyperplane of $d-1$ dimensions. Then all embeddings in $\boldsymbol{g_f}$ should be as far as possible from each other, so as to increase the discrimination between forgery domains, and weaken the similarities in forgery-irrelevant information. This optimization is formulated as:

$$L\left(\left\{\boldsymbol{g}_{f_i}\right\}_{i=1}^{N}\right) = \frac{1}{N}\sum_{i=1}^{N} \log \sum_{j=1}^{N} e^{\boldsymbol{g}_{f_i}^T \boldsymbol{g}_{f_j}/\tau} \tag{2}$$

where $\tau$ is a temperature parameter to control the scale of distribution [38]. We first obtain $\boldsymbol{g_r}$ by random initialization, and then take Eq. (1) as the constraint of Eq. (2) and solve this constrained optimization problem according to the Lagrangian multiplier method [4] to obtain $\boldsymbol{g_f}$. The solving process is detailed in Appendix 2. To guarantee that the equation is solvable, $d \geq N$.

Let $\theta_{ij}(i, j = 1, \cdots, N, i \neq j)$ represent the angle between $\boldsymbol{g_f}$. Adjusting $\theta_0$ can affect $\theta_{ij}$ and thus adjust the degree of separation between all embeddings. Note that $\theta_0$ is not as large as possible, because in the space with limited dimensions, the larger $\theta_0$, the smaller $\theta_{ij}$, and the separation $\theta_{ij}$ between forgery domains also needs to be maintained. With this well-designed guide-space, we can achieve both the separation of forgery domains and the compactness of the real domain. More importantly, this process is explicit and controllable, rather than implicit and uncontrolled learning.

### 3.2. Controllable optimization based on guide-space

In the optimization, let $\{(x_i, y_i, t_i)\}_{i=1}^{B}$ denote a batch of face images, where $y_i$ is the ground-truth label that marks whether the image is fake or not, i.e., $y_i \in \{0, 1\}$. $t_i$ refers to the domain label, i.e., $t_i \in [0, N]$, where for real faces, $t_i = 0$, and for fake faces, $t_i$ represents the forgery category label to which it belongs. The forgery detection model consists of a feature encoder $F(\cdot)$ followed by a binary classifier $h(\cdot)$.

Based on the pre-calculated guide embeddings in Sec. 3.1, we make the features of each domain close to their respective guide embeddings, so as to achieve the separation between real-forgery domains and between forgery-forgery domains. Let $v_i$ denote the feature of image $x_i$ extracted by $F(\cdot)$, $G$ denote the set of all guide embeddings, $G = \left\{\boldsymbol{g_r}, \boldsymbol{g_{f_i}} \mid i = 1, \cdots, N\right\}$, and the loss function can be formulated as:

$$L_{guide} = -\sum_{i=1}^{B} \lambda_i \log \frac{e^{v_i^T g_i^* / \tau}}{\sum_{v_j \in V \cup G} e^{v_i^T v_j / \tau}} \tag{3}$$

where $\lambda_i$ is the weight of the current data $x_i$ in the loss calculation relative to the data within a batch, and in general, the loss is the average of each data, i.e., $\lambda_i = 1/B$. $g_i^*$ is the guide embedding corresponding to $x_i$. $g_i^* = \begin{cases} g_r & \text{if } t_i = 0 \\ g_{f_j}\ (j = \Phi(t_i)) & \text{if } 1 \leq t_i \leq N \end{cases}$. $\Phi(\cdot)$ is the relation function between forgery domains and forgery guide embeddings. At each iteration, we compute the average feature for each forgery domain. According to the distance between each average feature and the guide embedding, we use Hungarian algorithm [20] to perform nearest-neighbor matching, and denote this matching relationship as $\Phi(\cdot)$. $V$ is a set that stores a large number of features (detailed in Sec. 3.4).

Besides, we also use the traditional binary cross-entropy loss as a basic optimization goal:

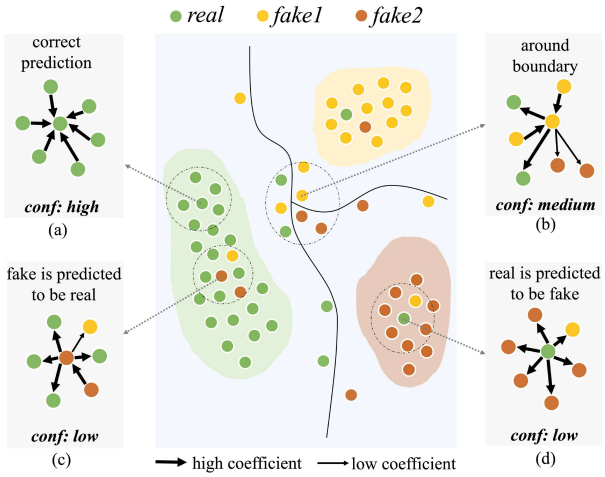


Figure 5. Four cases in the calculation of the confidence. For each center point, inward arrows indicate that classes are consistent, and outward arrows indicate class inconsistencies. The thickness of the arrow reflects the value of the coefficient $\mu$ in Eq. (5).

$$L_{ce} = -\sum_{i=1}^{B} \lambda_i \left(y_i \log p_i + (1 - y_i) \log (1 - p_i)\right) \tag{4}$$

where $p_i$ is the predicted score obtained by the binary classifier $h(\cdot)$. $\lambda_i$ is the weight consistent with Eq. (3).

### 3.3. Adjustment of decision boundary manifold

Furthermore, we design a module to adjust the decision boundary manifold (A-DBM). By focusing on some poorly performing samples within a batch, we strive to achieve the aggregation of features of the same domain, resulting in a better decision boundary manifold that conforms to the guidance of the guide-space.

Here, we define a metric called confidence to indicate the credibility of the model decision, which is calculated by the aggregation degree of the same domain features in the neighborhood of each feature point, and this process can be visualized in Figure 5. For a sample whose neighbors in the feature space belong to the same domain (Figure 5 (a)), the model prediction tends to be reliable and is assigned high confidence; while for the sample densely adjacent to other classes, it may be located in an area where another class is clustered, and tends to be mis-predicted as another class with high probability, corresponding to a low confidence (Figure 5 (c) and (d)). For the points around the decision boundary, the model decision has uncertainty and the confidence is at the median (Figure 5 (b)).

Specifically, for each image feature $v_i$, we calculate its similarity with each point in the feature set $V$, and then take the $k$ points with the highest similarity as its neighbors $K_i$ to obtain the adjacency relationship of the feature points. According to the aggregation degree of the same class of features within the neighbors, the confidence $c_i$ of each data

$x_i$ can be formulated as:

$$c_i = \frac{1}{|K_i|} \sum_{v_j \in K_i} \left( \mathbb{1}_{t_i = t_j} - \mu \cdot \mathbb{1}_{t_i \neq t_j} \right) \cdot \varepsilon_{ij} \tag{5}$$

where $\varepsilon_{ij}$ refers to the similarity between $v_i$ and $v_j$, and $\varepsilon_{ij} = \frac{1}{2}(1 + v_i v_j)$. $\mu$ is an adjustment coefficient. If $x_i$ is a real image, neighbors that are inconsistent with its class all belong to various fake categories with equal importance, i.e., $\mu = 1$. However, if $x_i$ is a forgery image, the class inconsistency includes two situations of real and other forgery domains. It is more important to separate it from the real data than to separate it from data of other forgery domains, so $\mu = 1$ for the former and $\mu = 0.5$ for the latter.

Based on the confidence, we can find samples that perform poorly according to the aggregation degree of features of the same domain. Then we assign higher attention (i.e., higher weight $\lambda_i$ in Eq. (3) and (4)) to these low-confidence samples to improve their clustering effect, resulting in a better decision boundary manifold. Given confidence $c_i$ ($i \in [B]$) of data within a batch, $\lambda_i$ can be formulated as:

$$\lambda_i = \mathrm{softmax}(-c_i) \tag{6}$$

### 3.4. Decoupling module for irrelevant information

In order for features to be more forgery-relevant, we design a decoupling module to alleviate the interference of irrelevant information on distinguishing different domains. Before training, we first mine the potential correlation between images of the given training data. We use a self-supervised model [25] pre-trained on face pictures to perform feature extraction on the training set. The self-supervised features have a certain degree of general representation ability for images. Then we cluster the features [9] to find correlations between the data, as shown in the decoupling module of Figure 3. If features of different forgery domains are gathered into the same cluster, it proves that they have strong similarities in forgery-irrelevant information and need to be focused on in training to be separated. Conversely, if the features of the same domain belong to different clusters, they should be pulled closer to make features more related to forgery.

Let $\rho_i$ denote the cluster label of each instance $x_i$, we first construct a sample set $V_i^+$ that needs to be pulled closer and a sample set $V_i^-$ that needs to be pushed away. $V_i^+ = \{v_j \in V \mid t_i = t_j, \rho_i \neq \rho_j\}$, $V_i^- = \{v_j \in V \mid t_i \neq t_j, \rho_i = \rho_j\}$. We randomly select $n^+$ and $n^-$ features from $V_i^+$ and $V_i^-$ respectively to participate in the calculation, and the corresponding feature set can be denoted as $\tilde{V}_i^+$ and $\tilde{V}_i^-$ ($\tilde{V}_i^+ \subset V_i^+$, $\tilde{V}_i^- \subset V_i^-$). The feature candidate set $V = \{(x_j, y_j, t_j)\}_{j=1}^{Q}$, which is a queue of dynamic accumulation of multiple batch features, with the current batch enqueued and the oldest batch dequeued. This $V$ with larger size $Q$ than the batch size allows better sampling in a broader and comprehensive selection scope.

On the one hand, we push away samples in $\tilde{V}_i^-$ that belong to different domains from $x_i$ but are clustered by irrelevant information. This separation helps to reduce the irrelevant information contained in the features. Based on KCL [18] loss, the pushing loss is denoted as:

$$L_{push} = \sum_{i=1}^{B} \lambda_i \cdot \frac{1}{1+n^-} \sum_{v_j^- \in \tilde{V}_i^-} \log \frac{e^{v_i^T v_j^- / \tau}}{\sum_{v_j \in V \cup G} e^{v_i^T v_j / \tau}} \tag{7}$$

On the other hand, we pull the samples in $\tilde{V}_i^+$ closer. This closeness due to the same domain rather than the similarity of irrelevant information implies an increase in the forgery-relevance of the features. The pulling loss is :

$$L_{pull} = -\sum_{i=1}^{B} \lambda_i \cdot \frac{1}{1+n^+} \sum_{v_j^+ \in \tilde{V}_i^+} \log \frac{e^{v_i^T v_j^+ / \tau}}{\sum_{v_j \in V \cup G} e^{v_i^T v_j / \tau}} \tag{8}$$

In summary, our overall loss can be formulated as:

$$L = \gamma_1 \cdot L_{guide} + \gamma_2 \cdot L_{ce} + \gamma_3 \cdot L_{pull} + \gamma_4 \cdot L_{push} \tag{9}$$

where $\lambda_i$ is calculated by Eq. (6), and $\gamma$ is the scale factor.

## 4. Experiments

### 4.1. Experimental settings

**Datasets:** We conduct experiments on three benchmark public forgery datasets: 1) FaceForensics++(FF++) [33] contains four forgery methods (i.e., Deepfakes (DF) [1], Face2Face (F2F) [42], FaceSwap (FS) [2], and NeuralTextures (NT) [41]) with three image qualities including raw, high quality (HQ) and low quality (LQ). 2) CelebDF [23] contains real videos of 59 celebrities and corresponding high-quality fake videos generated by the improved forgery techniques. 3) Deepfake Detection Challenge (DFDC) [11] is a more challenging dataset that comes with the competition, with many manipulation and perturbation methods.
**Metrics:** Following works in [47, 5, 24, 44, 26], we use Accuracy score (Acc) and Area Under the Receiver Operating Characteristic Curve (AUC) as the metrics to evaluate the performance. Between the two, we pay more attention to the AUC results, since the Acc values are affected by specific thresholds and data balance.
**Implementation Details:** In our experiments, we use EfficientNet-B4 (EN-B4) [40] as the backbone when not otherwise specified. In the guide-space, $\theta_0 = 120°$. For A-DBM, we adjust $\lambda_i$ from the 10-$th$ epoch, and before that, $\lambda_i = 1/B$. $k = |K_i| = 55$ in Eq. (5). In decoupling, the self-supervised model is trained under SimMIM [46] framework, and the number of clusters is 500. $n^+$ and $n^-$ are both 10. In Eq. (9), $\gamma_1 = 1$, $\gamma_2 = 0.5$, $\gamma_3 = 0.01$, and $\gamma_4 = 0.005$. The temperature parameter $\tau = 1$. During the training, the batchsize $B = 256$, and the size $Q$ of the set $V$ is 5120. The maximum number of epochs is 60. More details on hyper-parameters are shown in Appendix 3, and the analysis of computational cost introduced by the method is shown in Appendix 4.

| Methods | FF++ (HQ) | | FF++ (LQ) | |
|---|---|---|---|---|
| | Acc | AUC | Acc | AUC |
| Xception [32] | 95.04 | 96.30 | 84.11 | 92.50 |
| $F^3$-Net (Xception) [31] | 97.31 | 98.10 | 86.89 | 93.30 |
| EN-B4 [40] | 96.63 | 99.18 | 86.67 | 88.20 |
| MAT (EN-B4) [47] | 97.60 | 99.29 | 88.69 | 90.40 |
| SPSL [24] | 91.50 | 95.32 | 81.57 | 82.82 |
| RFM [44] | 95.69 | 98.79 | 87.06 | 89.83 |
| Local-relation [7] | 97.59 | 99.56 | 91.47 | 95.21 |
| RECCE [5] | 97.06 | 99.32 | 91.03 | 95.02 |
| CD-Net [36] | 98.75 | 99.90 | 88.12 | 95.20 |
| Ours | **99.24** | **99.95** | **92.76** | **96.85** |

Table 1. In-domain comparisons on FF++ dataset. Results contain Acc (%) and AUC (%) of high quality (HQ) and low quality (LQ).

| Methods | CelebDF | DFDC |
|---|---|---|
| Xception [32] | 66.91 | 67.93 |
| $F^3$-Net (Xception) [31] | 71.21 | 72.88 |
| EN-B4 [40] | 66.24 | 66.81 |
| MAT(EN-B4) [47] | 76.65 | 67.34 |
| Face X-ray [24] | 74.20 | 70.00 |
| RFM [44] | 67.64 | 68.01 |
| SRM [26] | 79.40 | 79.70 |
| Local-relation [7] | 78.26 | 76.53 |
| RECCE [5] | 77.39 | 76.75 |
| LTW [37] | 77.14 | 74.58 |
| DCL [38] | 82.30 | 76.71 |
| UIA-ViT [50] | 82.41 | 75.80 |
| Ours | **84.97** | **81.65** |

Table 2. Cross-domain comparisons of generalization based on AUC (%). We train the detection model on the HQ dataset of FF++ and then test it on CelebDF and DFDC.

### 4.2. In-Domain evaluations

We first verify the detection ability of our method against in-domain forgery methods (i.e., methods contained in the training set) on FF++ dataset. We train on both HQ and LQ image qualities using the four included forgery methods, and Table 1 lists the performance comparisons between ours and some current state-of-the-art methods.

It can be seen that our method achieves the best performance on HQ and LQ with AUC of 99.95% and 96.85%, respectively, confirming that our method is effective for both high-quality and low-quality data. On the HQ dataset, the AUC of our method is 1.16% higher than that of RFM which enlarges the model's attention by erasing sensitive areas. $F^3$-Net [31], SPSL [24], and CD-Net [36] consider the frequency domain information, and ours are all ahead of them. On the LQ dataset, our AUC is 6.45% higher than that of MAT [47] using the attention mechanism. Local-relation [7] fusing RGB and frequency domain information achieves the sub-optimal performance with the AUC of 95.21%, while ours is still 1.64% higher than that.

Although the above methods use attention, frequency information, local consistency, etc. to learn more generalized features, the supervision of binary classification makes the semantic and texture features still exist in the feature space. In contrast, our method increases the discrimination between domains through the guide-space, and decouples forgery-irrelevant information from features, so as to learn more forgery-relevant representations.

### 4.3. Cross-domain evaluations for generalization

To verify the generalization of our method, we train the models using four forgery methods on the HQ dataset of FF++, and then test the cross-domain generalization on CelebDF and DFDC. We compare with many state-of-the-art methods, such as DCL [38] using the contrastive learning (CL), $F^3$-Net [31], SRM [26], Local-relation [7] considering frequency domain information, UIA-ViT [50] based on the transformer, LTW [37] based on the meta-learning, and reconstruction-learning based RECCE [5]. The corresponding AUC results are shown in Table 2.

The AUC of our method on CelebDF and DFDC is 84.97% and 81.65%, respectively, outperforming other methods listed in Table 2. For CL-based methods, DCL focuses on the real-fake discrimination , achieving AUC of 82.30% and 76.71% on CelebDF and DFDC, while we further enhance the discrimination of different domains (real-fake, fake-fake), and AUCs are 2.67% and 4.94% higher than it, respectively. For frequency-based methods, ours is 5.57% ahead of SRM [26] on CelebDF. For recent transformer-based UIA-ViT, our AUC leads by 2.56% and 5.85% on Celeb-DF and DFDC, respectively. Although RECCE [5] also learns the common compact representations of real faces through reconstruction learning, its AUCs on the two datasets are 77.39% and 76.75%, which are lower than ours. Unlike implicit learning in RECCE, we explicitly control how compact the real representation is by controlling the angle $\theta_0$ between the real and forgery embeddings in the guide-space. So coupled with the efforts that we also focus on the discrimination between different forgery domains to capture more forgery-related traces, we can achieve better performance.

### 4.4. Ablation study

In this section, we conduct detailed studies of each module involved in the method. The evaluation of generalization follows two settings: 1) Cross-test within FF++: training with three methods within FF++ and testing on the remaining one; 2) From FF++ to others: training on the four forgery methods of FF++ and testing on CelebDF and DFDC.

**Methods to distinguish different forgery domains:** We first analyze methods of enhancing the forgery domain discrimination to verify our superiority based on the guide-space. The training result based on binary cross-entropy

| Train Set | F2F FS NT | | | | DF FS NT | | | | FF++ (HQ) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test Set | DF (HQ) | | DF (LQ) | | F2F (HQ) | | F2F (LQ) | | CelebDF | | DFDC | |
| | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC | Acc | AUC |
| $L_{ce-2}$ | 91.97 | 92.48 | 91.46 | 96.71 | 84.44 | 91.17 | **91.37** | 96.49 | 62.93 | 66.24 | 62.16 | 66.81 |
| $L_{ce-(1+N)}$ | 93.83 | 95.59 | 92.34 | 97.32 | 83.91 | 91.15 | 91.08 | 96.42 | 65.06 | 67.74 | 65.16 | 68.58 |
| $L_{guide}$ | **94.21** | **95.76** | **93.96** | **97.79** | **85.54** | **92.33** | 90.42 | **96.87** | **66.02** | **70.72** | **67.67** | **71.47** |
| w/o $L_{guide}$ | 93.07 | 96.91 | 93.62 | 97.64 | 87.85 | 93.56 | 95.72 | 98.67 | 68.59 | 76.42 | 69.37 | 75.12 |
| w/o $L_{pull}$&$L_{push}$ | 95.82 | 98.15 | 95.73 | 98.59 | 90.26 | 94.19 | 95.90 | 99.08 | 71.52 | 79.13 | 72.85 | 78.04 |
| w/o $L_{pull}$ | 97.12 | 98.87 | 96.62 | 99.11 | 92.65 | 95.82 | 96.17 | 99.32 | 71.63 | 80.25 | 73.09 | 78.98 |
| w/o $L_{push}$ | 97.46 | 99.39 | 97.71 | 99.57 | 93.14 | 97.50 | 96.91 | 99.49 | 71.94 | 81.78 | 73.55 | 79.26 |
| w/o A-DBM | 96.03 | 97.40 | 96.47 | 98.79 | 91.28 | 96.35 | 95.87 | 99.02 | 72.12 | 78.92 | 70.23 | 77.15 |
| ours | **98.92** | **99.81** | **98.72** | **99.89** | **95.76** | **98.92** | **97.96** | **99.68** | **73.19** | **84.97** | **74.83** | **81.65** |

Table 3. Ablation results in two settings: 1) Cross-test within FF++ (left); 2) From FF++ to others (right). The upper part compares the three losses that increase the discrimination of different domains; the lower part shows the performance after removing each module of the method.

($L_{ce-2}$) loss is an experimental baseline. On this basis, we compare the performance of using multi-class cross-entropy ($L_{ce-(1+N)}$) loss, where the number of classes is the number of forgery domains $N$ plus 1. The performance comparisons of $L_{ce-2}$, $L_{ce-(1+N)}$, and our $L_{guide}$ under two experimental settings are shown in the upper part of Table 3. For the cross-test within FF++, we experiment on both HQ and LQ datasets. Due to space limitations, we here show the results of DF and F2F as the test set, and the remaining results of FS and NT are shown in Appendix 5.1.

It can be seen that the AUC of our guide-method $L_{guide}$ achieves optimal performance among the three losses at all of these settings. Especially on CelebDF and DFDC, AUCs are 4.48% and 4.66% higher than $L_{ce-2}$, and 2.98% and 2.89% higher than $L_{ce-(1+N)}$. For $L_{ce-(1+N)}$, it has improved performance over $L_{ce-2}$ in most cases, but in some cases, it is the opposite, for example, Acc and AUC on F2F(HQ) are 0.53% and 0.02% lower than $L_{ce-2}$. This shows that simply uncontrolled separation of several forgery domains in $L_{ce-(1+N)}$ is sometimes not feasible, because it treats the real and forgery domain as equal classes, so that the distances between real-forgery and between forgery-forgery are equal. However, we expect the separation of real-forgery to be much greater than forgery-forgery to achieve the compactness of the real domain representation. Our guide method can achieve this by actively controlling the angle between the real and forgery guide embeddings, so it can achieve the optimal performance.

**Importance of different modules:** We mainly study the guide embedding ($L_{guide}$), the decoupling module ($L_{pull}$, $L_{push}$), and the boundary adjustment (A-DBM) included in the method, and examine the importance by removing the corresponding module from the overall method. The results are shown in the lower half of Table 3, and the remaining results of FS and NT are shown in Appendix 5.2.

Overall, removing the guide method (w/o $L_{guide}$) has the greatest impact, e.g., the AUC on CelebDF is reduced from 84.97% to 76.42%. Although the decoupling module can also achieve the separation of different domains and the closeness of the same domain, it lacks predefined guide vectors and cannot actively control the degree of separation between domains. The performance also drops significantly when the decoupling module is removed (w/o $L_{pull}$&$L_{push}$), demonstrating the importance of alleviating the association of irrelevant information for features. Based on the results, removing the pulling set (w/o $L_{pull}$) has a greater negative impact than removing the pushing set (w/o $L_{push}$), for example, on CelebDF, AUCs decrease by 4.72% and 3.19%, respectively. The boundary module works well in all experimental settings, especially with 6.05% and 4.5% increase of AUC on CelebDF and DFDC after using it. In addition, under the left setting (1) of Table 3, the performance comparison of our method with other SOTA methods is shown in Appendix 5.3.

| Backbone | CelebDF | | DFDC | |
|---|---|---|---|---|
| | Acc | AUC | Acc | AUC |
| Xception | 64.09 | 66.91 | 62.16 | 67.93 |
| Xception+Ours | 69.31 | 76.46 | 66.29 | 75.21 |
| Resnet50 | 62.47 | 67.08 | 64.16 | 67.68 |
| Resnet50+Ours | 71.92 | 77.05 | 70.69 | 74.03 |
| DPN68 | 64.86 | 70.78 | 64.16 | 67.72 |
| DPN68+Ours | 72.73 | 80.09 | 68.98 | 77.25 |
| VGG19 | 67.18 | 71.31 | 68.17 | 72.56 |
| VGG19+Ours | 74.91 | 81.89 | 71.54 | 78.75 |

Table 4. Generalization when using other backbones. Models are trained on FF++(HQ) and tested on CelebDF and DFDC.

**Generalization when using other backbones:** Besides EfficientNet-B4 [40], our method can also be used in other backbone networks. Table 4 lists the performance of our method under Xception [32], Resnet50 [15], DPN68 [8] and

| $\theta_0$ | $\theta_{ij}$ | CelebDF | | DFDC | |
|---|---|---|---|---|---|
| | | Acc | AUC | Acc | AUC |
| 90° | 109° | 69.22 | 74.95 | 66.34 | 74.82 |
| 100° | 107° | 70.19 | 77.86 | 71.74 | 76.95 |
| 110° | 100° | 72.58 | 81.07 | 73.91 | 79.23 |
| 120° | 90° | **73.19** | **84.97** | **74.83** | **81.65** |
| 130° | 78° | 71.32 | 80.83 | 73.56 | 80.19 |
| 140° | 63° | 70.78 | 79.05 | 71.29 | 78.46 |
| 150° | 48° | 70.20 | 78.52 | 70.96 | 77.87 |

Table 5. The value of $\theta_{ij}(1 \leq i, j \leq N)$ and the corresponding generalization performance when $\theta_0$ takes different values.

VGG19 [35]. The models are trained on FF++ (HQ) and tested on CelebDF and DFDC. Compared with the original training, the performance on two datasets has improved after using our training method, e.g., AUCs of the four models on CelebDF are improved by 9.55%, 9.97%, 9.31%, and 10.58%, respectively. This demonstrates that our method has good adaptability and can be combined with various backbone networks to achieve better performance.

### 4.5. Analysis and visualization

**Predefined $\theta_0$ in guide-space:** When constructing the guide-space, we pre-define the angle $\theta_0$ of real-forgery domains, and maximize the forgery-forgery domain separation $\theta_{ij}(1 \leq i, j \leq N)$. $\theta_{ij}$ can be equal when $d \geq N$. However, in the limited feature space, $\theta_0$ and $\theta_{ij}$ have a trade-off, i.e., a large $\theta_0$ means that the corresponding $\theta_{ij}$ will be small. We study the effect of different $\theta_0$, and the corresponding $\theta_{ij}$ and the generalization results are listed in Table 5.

Ideally, we expect $\theta_0 > \theta_{ij}$, and $\theta_{ij}$ is as large as possible. In Table 5, the optimum performance is achieved when $\theta_0 = 120°$, and the corresponding $\theta_{ij} = 90°$. $\theta_0$ is larger than $\theta_{ij}$, and the forgery-forgery is orthogonal, realizing the separation of each other. When we decrease $\theta_0$ to 90°, $\theta_{ij}$ increases to 109°, and the performance drops significantly, e.g., the AUC of CelebDF decreases from 84.97% to 74.95%. At this time, although the forgery-forgery distinction is enhanced, the real and forgery domains are too close, which will limit the generalization. Conversely, when we increase $\theta_0$ to 150° , the real-forgery distance is increased, but the forgery-forgery $\theta_{ij}$ is reduced to 48°. As a result, the discrimination between forgery domains is weakened, and the AUC on CelebDF drops to 78.52%.

**Visualizations of t-sne:** In Figure 7, we visualize the feature space using t-sne [43], comparing the effect of binary cross-entropy (CE-2) with our method. For CE-2, the points of the forgery domains are messily mixed together without being distinguished by forgery domains. Instead, the features of ours are clustered according to the domain type, indicating that the learned features are more forgery-related.

**Visualizations of heatmaps:** Identity is one of the forgery-irrelevant similarities existed between different domains. In Figure 6, we compare the heatmaps for images of different forgery domains with the same identity when using CE-2 and our method. When using CE-2, different domains are treated as the same fake, resulting in the model focusing on similar regions. As shown in the second line of Figure 6, the concern areas are concentrated in the middle of the face rather than their respective forgery traces, indicating that the model is still perturbed by similarities of the same identity. However, in the third line, our method focuses on respective forgery traces of different forgeries, which shows that our method mitigates the interference of these similarities and learns better forgery-related information. More heatmaps are shown in Appendix 6.

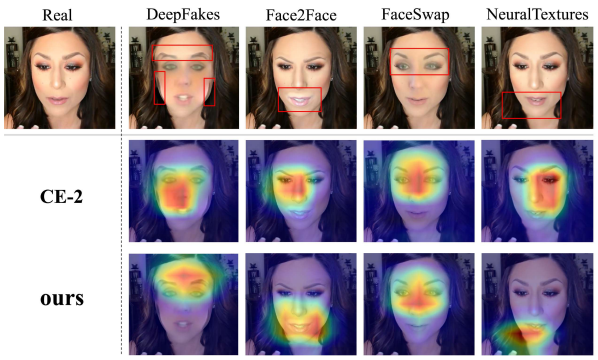


Figure 6. The heatmap comparisons of binary cross-entropy (CE-2) and our method. Forgery artifacts are marked in red frames.


(a) Traditional training with CE-2 (a) Training with our method

Figure 7. The t-sne comparisons of binary cross-entropy (CE-2) and our method.

## 5. Conclusion

In this paper, we proposed a novel guide-space based framework to improve the generalization of face forgery detectors. The well-designed guide-space can achieve separations of both real-forgery domains and forgery-forgery domains in a controllable manner, so as to capture more forgery-related information and ensure a large distance between real and fake representations simultaneously. Furthermore, we used a decoupling module to reduce the interference of forgery-irrelevant inter-domain correlations for domain discrimination. In addition, we designed a decision boundary adjustment module to make the features better follow the guidance of the guide-space. Extensive cross-domain experiments demonstrate the better generalization of our method.

## References


[1] Deepfakes. https://github.com/deepfakes/faceswap. Accessed 2022-10-29.

[2] Faceswap. https://github.com/MarekKowalski/FaceSwap. Accessed 2022-10-29.

[3] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[4] Brian Beavis and Ian Dobbs. *Optimisation and stability theory for economic analysis*. Cambridge university press, 1990.

[5] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022.

[6] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18710–18719, 2022.

[7] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1081–1088, 2021.

[8] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. *Advances in neural information processing systems*, 30, 2017.

[9] Yingjie Chen, Huasong Zhong, Chong Chen, Chen Shen, Jianqiang Huang, Tao Wang, Yun Liang, and Qianru Sun. On mitigating hard clusters for face clustering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pages 529–544. Springer, 2022.

[10] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.

[11] 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.

[12] Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, and Jian Weng. Learning second order local anomaly for general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20270–20280, 2022.

[13] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3404–3413, 2021.

[14] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33:9995–10006, 2020.

[17] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: Robust deepfake detection using frequency-level perturbations. 2022.

[18] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020.

[19] Minha Kim, Shahroz Tariq, and Simon S Woo. Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1001–1012, 2021.

[20] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[21] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6458–6467, 2021.

[22] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[23] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020.

[24] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021.

[25] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[26] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021.

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

In *European conference on computer vision*, pages 667–684. Springer, 2020.

[28] Aakash Varma Nadimpalli and Ajita Rattani. On improving cross-dataset generalization of deepfake detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 91–99, 2022.

[29] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2019.

[30] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311. IEEE, 2019.

[31] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

[32] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[33] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[34] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Luchuan Song, Zheng Fang, Xiaodan Li, Xiaoyi Dong, Zhenchao Jin, Yuefeng Chen, and Siwei Lyu. Adaptive face forgery detection in cross domain. In *European Conference on Computer Vision*, pages 467–484. Springer, 2022.

[37] 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 *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2638–2646, 2021.

[38] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2316–2324, 2022.

[39] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3609–3618, 2021.

[40] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[41] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[42] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[44] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14923–14932, 2021.

[45] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2022.

[46] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[47] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.

[48] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021.

[49] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1831–1839. IEEE, 2017.

[50] 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. *arXiv preprint arXiv:2210.12752*, 2022.