

# *Trident*: Detecting Face Forgeries with Domain-adversarial Triplet Learning

Mustafa Hakan Kara, Aysegul Dundar, Uğur Gündükbay

**Abstract**—As face forgeries generated by deep neural networks become increasingly sophisticated, detecting face manipulations in digital media has posed a significant challenge, underscoring the importance of maintaining digital media integrity and combating visual disinformation. Current detection models, predominantly based on supervised training with domain-specific data, often falter against forgeries generated by unencountered techniques. In response to this challenge, we introduce *Trident*, a face forgery detection framework that employs triplet learning with a Siamese network architecture for enhanced adaptability across diverse forgery methods. *Trident* is trained on curated triplets to isolate nuanced differences of forgeries, capturing fine-grained features that distinguish pristine samples from manipulated ones while controlling for other variables. To further enhance generalizability, we incorporate domain-adversarial training with a forgery discriminator. This adversarial component guides our embedding model towards forgery-agnostic representations, improving its robustness to unseen manipulations. In addition, we prevent gradient flow from the classifier head to the embedding model, avoiding overfitting induced by artifacts peculiar to certain forgeries. Comprehensive evaluations across multiple benchmarks and ablation studies demonstrate the effectiveness of our framework. We will release our code in a GitHub repository.

**Index Terms**—Deepfake detection, triplet learning, domain-adversarial training, generalizable face forgery detection, domain generalization.

## I. INTRODUCTION

Advances in deep-generative networks, particularly Generative Adversarial Networks (GANs) [1] and, more recently, diffusion models [2], have led to increasingly sophisticated visual forgeries, posing significant challenges to digital media integrity and public discourse. As manipulated images become prevalent, traditional supervised approaches have shown limitations in:

- generalization across diverse manipulation techniques and
- detection of previously unseen forgery methods [3].

To address these challenges, we present *Trident* (Triplet learning-based Deepfake detection Network), a triplet learning-based approach designed to generalize across a wide range of forgery techniques, including those not encountered during training. Our method builds upon triplet loss techniques originally introduced for face recognition tasks [4] but adapts them to the more complex objective of detecting forged imagery. By carefully curating triplets that preserve the person ID and scene context, *Trident* encourages the model to isolate

subtle, forgery-specific features rather than relying on incidental factors like background or identity. This design ensures that the learned embeddings capture the intrinsic differences between genuine and manipulated samples, thereby improving generalization and the detection process’s robustness.

*Trident* also leverages domain-adversarial training [5] with a forgery discriminator, which attempts to classify the forgery category given an embedding. This adversarial interplay pushes the embedding generator to produce representations agnostic to specific forgery types, focusing instead on the fundamental artifacts that distinguish real from fake. Such an adversarial mechanism reduces overfitting to known manipulation methods, resulting in discriminative embeddings even when confronted with novel, unseen forgeries.

In addition, we detach Siamese network embeddings before passing them to the classifier head, ensuring the classifier’s gradients do not influence the embedding network during backpropagation. This approach maintains the integrity of the embedding space learned through triplet loss, allowing it to focus more on capturing generalizable forgery artifacts. For our transformer-based variant, we employ BitFit [6], a parameter-efficient fine-tuning approach that prevents catastrophic forgetting while enabling efficient adaptation to the forgery detection task.

Our main contributions to face forgery detection are

- a novel controlled triplet learning formulation that disentangles manipulation cues from identity and scene-specific information by maintaining consistent identity and temporal alignment,
- the introduction of an adversarially trained forgery discriminator, guiding the embeddings toward forgery-agnostic representations that remain effective against unseen manipulations,
- detaching backbone embeddings from the binary classifier head to further increase generalization, and
- efficient bias-only fine-tuning to avoid catastrophic forgetting while adapting to forgery detection with minimal parameter updates.

Extensive experiments demonstrate that the *Trident* framework achieves competitive performance on challenging benchmarks while exhibiting more adaptability than traditional supervised approaches. Our findings highlight the potential of combining triplet learning, domain-adversarial training, and parameter-efficient fine-tuning to detect the evolving landscape of face forgeries, contributing to more robust methods for preserving digital media authenticity.

M. H. Kara, A. Dundar, and U. Gündükbay are with the Department of Computer Engineering, Bilkent University, Ankara, Turkey.  
E-mail: hakan.kara@bilkent.edu.tr, {adundar, gudukbay}@cs.bilkent.edu.tr  
Corresponding author: U. Gündükbay.

## II. RELATED WORK

### A. Spatial-Domain Methods

Face-forgery detection methods have evolved significantly to address increasingly sophisticated manipulation techniques [7]. Early approaches relied on detecting specific artifacts left by generation algorithms. Xception [8] and EfficientNet [9] emerged as strong baseline models due to their robust performance on benchmark datasets [10]. *Face X-Ray* [11] focused on blending artifacts in face-swapping operations, while Chai *et al.* [12] examined differences between camera-captured and algorithm-manipulated images.

Recent spatial-domain detectors integrate explicit artifact reasoning and finer attention. Li *et al.* [13] disentangle forgery artifacts from identity information via adversarial learning, Nguyen *et al.* [14] employ localized artifact attention to remain quality-agnostic, and Masked-relation learning [15] models region-to-region relations as a graph, propagating forgery cues globally. Dagar and Vishwakarma [16] proposed a dual-branch architecture that fuses noise cues with hierarchical ConvNeXt features, attaining superior manipulation-localization accuracy on both shallowfake and deepfake datasets. Yadav and Vishwakarma [17] proposed Face-NeSt, which leverages adaptively weighted multi-scale attentional features and underscores the importance of balanced scale fusion. Long *et al.* [18] introduced LGDF-Net, a dual-branch fusion network that separately processes local artifact and global texture features through specialized compression and expansion modules.

### B. Temporal and Frequency Domain Methods

Temporal detectors exploit inconsistencies across video frames. Yang *et al.* [19] flagged abnormal head-pose dynamics, while LipForensics [20] tracked mouth-movement coherence. Gu *et al.* [21] proposed Spatio-Temporal Inconsistency Learning, combining spatial cues with a temporal stream; Amerini *et al.* [22] leveraged optical-flow CNNs to uncover compression-robust artefacts. Choi *et al.* [23] detected style-latent incoherence between frames, and Cheng *et al.* [24] introduced a cross-modal strategy that measures voice–face homogeneity, exposing identity mismatches characteristic of face-swap videos. Yu *et al.* [25] magnify subtle spatio-temporal anomalies via multi-timescale views, while Lu *et al.* [26] use long-distance attention to capture global spatial–temporal cues. Xu *et al.* [27] introduced TALL, which transforms video clips into pre-defined layouts to preserve spatial-temporal dependencies while being computationally efficient. Li *et al.* [28] leveraged robust facial landmarks with spatial and temporal rotation angles to achieve compression-resistant detection.

Frequency-domain methods analyze spectral artifacts introduced during synthesis. Qian *et al.* [29] proposed F<sup>3</sup>-Net to mine frequency-aware clues; Li *et al.* [30] learned discriminative features across bands; Liu *et al.* [31] exploited phase discrepancies. Frank *et al.* [32] linked GAN up-sampling to high-frequency artefacts. Tan *et al.* revisited up-sampling operations by modelling neighbouring-pixel relationships for open-world detection [33]; their follow-up FreqNet [34] enforces frequency-aware learning to obtain source-agnostic detectors.

### C. Vision Transformer-based Forgery Detection

Vision Transformers (ViTs) have emerged as powerful architectures for deepfake detection due to their ability to capture long-range dependencies and fine-grained features [35]. ISTVT [36] combines spatial and temporal attention, while M2TR [37] analyzes images at multiple scales through parallel transformer branches. ICT [38] verifies identity consistency, and UIA-ViT [39] models feature distributions without pixel-level masks. Khan *et al.* [40] fuse UV texture maps with transformers for pose-invariant detection, and GenConViT [41] combines generative priors. Forgery-Aware Adaptive Learning ViT (FAL-ViT) [42] further enhances generalization by adapting a transformer backbone to forgery cues during continual domain shifts. Dagar and Vishwakarma introduced Tex-ViT [43], which augments a ResNet backbone with a texture-aware module and a dual-branch cross-attention Vision Transformer.

### D. Generalization-Focused Methods

A critical challenge in deepfake detection is achieving robustness across unseen manipulations and real-world conditions. Yu *et al.* [44] attributed synthetic images to source GANs by learning fingerprints. Ni *et al.* [45] enforced representation consistency via *CORE*. Zhao *et al.* [46] introduced self-consistency learning.

Huang *et al.* [47] disentangled identity information, Yan *et al.* [48] decomposed images into forgery-relevant partitions, and Ojha *et al.* [49] used a frozen encoder plus classical classifiers. SBIs [50] employ self-blending, while RECCE [51] reconstructs authentic faces. Zhang *et al.* [52] mitigate catastrophic forgetting via expansive learning, and cross-modal strategies such as PVASS-MDD [53] and MCL [54] leverage alignment and contrastive objectives for improved generalization.

### E. Contrastive and Metric Learning Approaches

Contrastive and metric learning enhance generalizability by enforcing feature discrimination. Fung *et al.* [55] proposed DeepfakeUCL, an unsupervised contrastive scheme. Xu *et al.* [56] developed supervised contrastive learning, while Gu *et al.* [57] introduced hierarchical video contrast. Liang *et al.* [58] used depth-guided triplets, and Kumar *et al.* [59] applied triplet networks in high-compression scenarios.

### F. Domain-Adversarial Training

Domain-adversarial training learns domain-invariant representations. Ganin and Lempitsky [5] introduced the Gradient Reversal Layer for unsupervised adaptation. EANN [60] applied similar ideas to fake news detection. Our work builds on these principles, combining adversarial and triplet objectives to realize forgery-agnostic representations.

## III. METHOD

Existing deepfake detection methods often struggle with generalization to unseen forgery techniques due to their

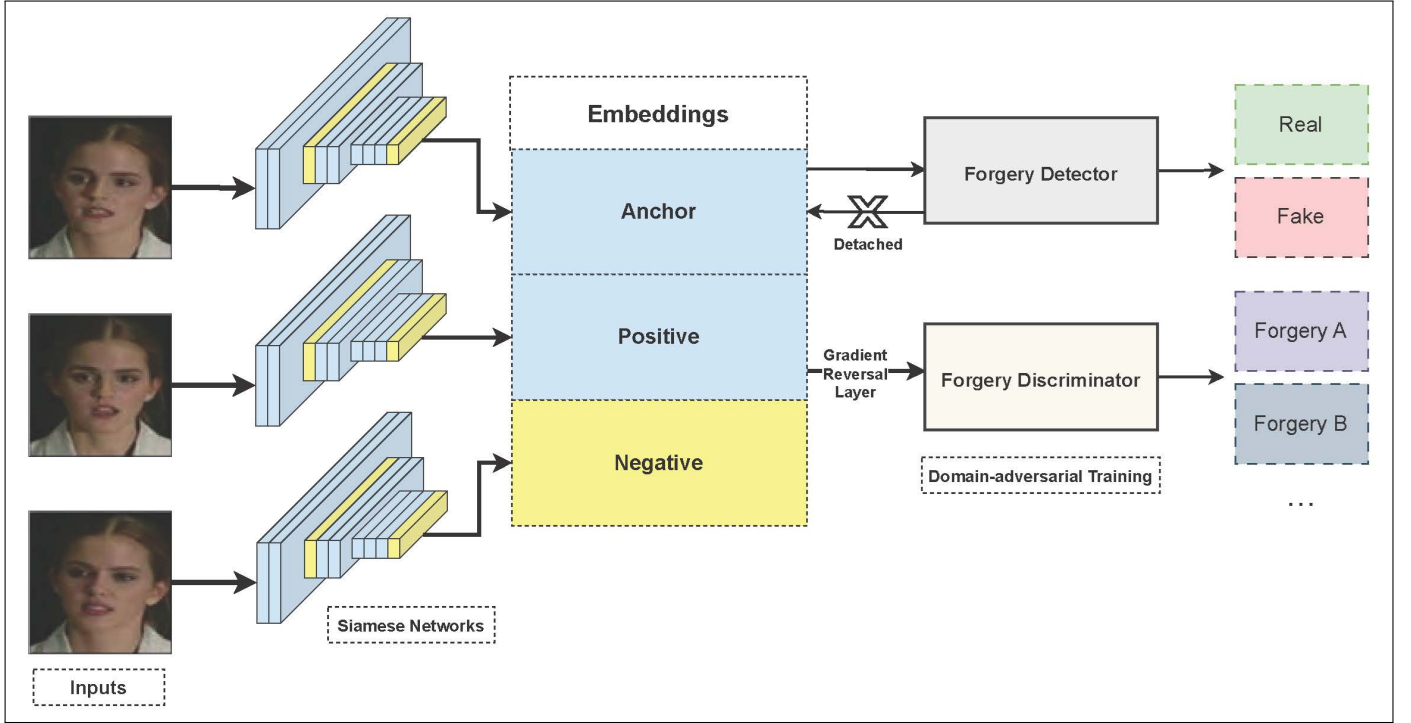


Fig. 1: Overview of our high-level architecture. We leverage a Siamese network with a triplet learning setup to generate embeddings. The forgery detector performs real-fake classification, while the forgery discriminator performs forgery-type classification with a Gradient Reversal Layer to enhance generalization.

reliance on forgery-specific artifacts. Unlike previous approaches, we apply triplet learning to deepfake detection in a controlled setting. By selecting triplets that maintain a consistent person ID and scene, we isolate forgery-specific features, prompting the model to focus on subtle and discriminative artifacts introduced by the forgery process rather than incidental variations in identity or background.

We also integrate a domain-adversarial training component to drive the network towards learning forgery-agnostic features, employing a forgery discriminator to classify specific manipulation methods. The embedding generator seeks to produce representations indistinguishable to this discriminator, thus pushing the model to rely on universal markers of manipulation rather than forgery-type-specific artifacts. This adversarial interplay leads to more generalized embeddings that better handle emerging and previously unseen forgery techniques.

#### A. Triplet Learning Framework

*Trident* employs a triplet learning setting [4], [61], building upon the foundations of contrastive learning and Siamese networks. Contrastive learning, originally proposed for face verification tasks [62], aims to distinguish samples by teaching the model to pull similar items closer and push dissimilar ones apart in the feature space. It employs contrastive loss on pairs of data points, penalizing large distances between similar and small distances between dissimilar pairs [63].

Siamese networks, introduced by [64], involve feeding pairs of samples into identical neural networks with shared

weights. This architecture typically measures the similarity of two inputs regarding learned feature representations. As described in [4], triplet learning extends these concepts by introducing a third element, creating anchor-positive-negative sample triplets.

This approach aims to bring the anchor and positive samples closer together while pushing the anchor and negative samples farther apart in the feature space. We leverage this triplet learning setting in *Trident*, training a projection head with triplets consisting of two real samples (anchor and positive) and one fake sample (negative).

We leverage this idea to learn the fine-grained features that distinguish genuine samples from forgeries while controlling for other variables such as scene and person ID. This formulation allows us to capture nuances and forces the model to discern subtle differences and similarities, potentially improving our model’s ability to generalize to previously unseen forgeries.

#### B. Model Architecture

Our model architecture, shown in Figure 1, involves a Siamese network structure with a triplet learning setup to distinguish real samples from fake ones. By curating triplets with consistent person ID and scene, we apply triplet learning to deepfake detection in a controlled setting. This controlled triplet learning setting isolates forgery-specific features, allowing the model to focus on finer details and more discriminative features related to the forgery process rather than incidental variations in identity or background.

The architecture comprises four key components visible in Figure 1: (1) a *Siamese network* with shared weights that processes triplet inputs and generates embeddings, (2) a *forgery detector* that performs binary real-fake classification on these embeddings, (3) a *forgery discriminator* that attempts to classify forgery types and is connected via a Gradient Reversal Layer (GRL) to promote domain-adversarial training, and (4) a *classifier head detachment mechanism* that isolates the binary classification from embedding generation to preserve generalization capability. This section provides an in-depth explanation of each component in our architecture.

1) *Backbone Architecture and Fine-tuning*: We fine-tuned two backbone variants for comprehensive evaluation: EfficientNet-B4 [9], a CNN-based architecture commonly preferred across face forgery detection methods, and CLIP ViT-L/14 [65], OpenAI’s transformer-based foundational model pretrained on large-scale visual and textual data.

For the ViT backbone, we employed BitFit [6], a parameter-efficient fine-tuning (PEFT) method that updates only the bias parameters while keeping all weights frozen. BitFit prevents catastrophic forgetting of learned abstract representations from pretraining while enabling efficient adaptation to the forgery detection task. This approach significantly reduces the number of trainable parameters while preserving the model’s generalization capability. Table I compares the two backbones we fine-tuned during our experiments.

TABLE I: Backbone Comparison

Backbone	EfficientNet-B4 [9]	CLIP ViT-L/14 [65]
Architecture	CNN	Transformer
Fine-tuning	Full	BitFit (bias-only)
Embedding Dimension	1792	1024
Total Parameters	19.3M	304M
Trainable Parameters	19.3M	0.3M

2) *Siamese Network and Embedding Generation*: We process each triplet of inputs through three parallel Siamese networks with shared weights, generating embeddings that capture fine-grained, distinguishing features between genuine and forged samples. Our triplet loss structure encourages the model to minimize the distance between the anchor and positive samples while maximizing the distance between the anchor and negative samples in feature space. By designing embeddings like this, we enable the model to learn representations robust to variations such as person ID or scene context, improving its generalization ability.

3) *Forgery Detector*: We pass these embeddings into the *Forgery Detector*, a binary classifier determining whether a sample is real or fake. By dedicating this head to binary classification, we allow the model to focus on separating real samples from forgeries without directly identifying the forgery type. This separation improves generalization since the *Forgery Detector* concentrates exclusively on the real versus fake distinction, enabling the model to adapt more effectively to unseen forgery types.

4) *Domain-adversarial Training Setup*: Our framework employs domain-adversarial training to learn forgery-agnostic representations that generalize across different manipulation

techniques. The core principle involves a minimax game between the embedding generator and a forgery discriminator. The embedding generator aims to produce informative features for real-fake classification. However, it is uninformative for forgery-type classification, while the forgery discriminator attempts to classify the specific forgery technique from these embeddings.

This adversarial setup encourages the embedding generator to focus on universal manipulation artifacts rather than technique-specific signatures. The forgery discriminator consists of multiple fully connected layers with an output layer producing logits for each forgery category (Deepfakes, Face2Face, FaceSwap, NeuralTextures, and real). Through this adversarial interaction, the model learns to capture broad real-versus-fake distinctions while avoiding overfitting to particular forgery methods.

5) *Gradient Reversal Layer*: The Gradient Reversal Layer (GRL) [5] implements the adversarial training mechanism by reversing gradients during backpropagation. The GRL is defined such that it acts as an identity mapping in the forward pass but reverses and scales the gradient by a factor  $\lambda$  in the backward pass. Let  $x$  be the input embeddings and  $\lambda > 0$ :

$$\text{Forward: } G(x) = x \quad (1)$$

During the backward pass, the gradient is multiplied by  $-\lambda$ :

$$\text{Backward: } \frac{\partial G}{\partial x} = -\lambda I \quad (2)$$

where  $I$  is the identity matrix. This scaling factor  $\lambda$  allows fine-grained control over the adversarial signal.

The GRL is placed between the embedding generator and the forgery discriminator. It acts as a pass-through layer during forward propagation but reverses the gradient from the forgery discriminator’s multi-class cross-entropy loss during backpropagation. This reversal encourages the embedding generator to produce feature representations that confuse the forgery discriminator, thereby learning more generalizable, forgery-invariant embeddings. The domain-adversarial training with the GRL aligns the embedding generator’s objective with producing domain-invariant features, preventing the model from overfitting to known forgery types.

6) *Classifier Head Detachment*: We isolate the *Forgery Detector* (real-fake classification head) from the embedding generation process to further enhance generalization. In practice, we block backpropagation from the forgery detector into the shared embedding layers. By preventing gradient flow from the binary classification task back into the embedding space, we ensure that the embeddings remain focused on forgery-related features rather than overfitting to the particularities of the classification head’s decision boundary. This separation helps maintain a more stable and general-purpose embedding space, improving the model’s adaptability to previously unseen forgeries.

### C. Objective Functions

The training objective in *Trident* is optimized using a hybrid loss function that combines *Binary Cross-Entropy (BCE) loss*, *Triplet loss*, and *Forgery Discriminator loss*. This composite



loss is designed to enhance class discrimination (via BCE), improve feature embedding separation (via Triplet loss), and promote forgery-agnostic embedding generation (via adversarial training through GRL).

The total loss  $\mathcal{L}_{\text{total}}$  is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \alpha \cdot \mathcal{L}_{\text{triplet}} + \beta \cdot \mathcal{L}_{\text{forgery}} \quad (3)$$

where  $\alpha$  and  $\beta$  are weighting factors that balance the contributions of the individual loss components.

1) *Binary Cross-Entropy Loss (BCE)*: The Binary Cross-Entropy loss  $\mathcal{L}_{\text{BCE}}$  is defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N (y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)) \quad (4)$$

where  $N$  is the number of samples,  $y_i$  is the ground truth label, and  $\hat{y}_i$  is the predicted probability of the sample being a forgery. This loss encourages accurate classification by minimizing the error between the predicted and actual labels.

2) *Triplet Loss*: The Triplet loss  $\mathcal{L}_{\text{triplet}}$  is formulated as:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{N} \sum_{i=1}^N [\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 - m]_+ \quad (5)$$

where  $x_i^a$ ,  $x_i^p$ , and  $x_i^n$  represent the anchor, positive, and negative samples, respectively, and  $m$  is the margin that enforces a minimum separation between positive and negative pairs. This loss function brings the anchor and positive embeddings closer together while pushing the anchor and negative embeddings farther apart, thereby enhancing feature distinctiveness.

3) *Forgery Discriminator Loss*: The forgery discriminator loss  $\mathcal{L}_{\text{forgery}}$  employs multi-class cross-entropy to classify forgery types:

$$\mathcal{L}_{\text{forgery}} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^K y_{i,j} \log(\hat{p}_{i,j}) \quad (6)$$

where  $y_{i,j}$  is the one-hot encoded ground truth for the  $j$ -th forgery category,  $\hat{p}_{i,j}$  is the predicted probability of the  $i$ -th embedding belonging to the  $j$ -th forgery category, and  $K$  is the number of forgery categories. During backpropagation, the GRL reverses the gradients from this loss, encouraging the embedding generator to produce forgery-agnostic representations.

#### D. Triplet Formation

Our controlled triplet formation strategy represents a key methodological contribution that distinguishes our approach from standard triplet learning-based methods. We constructed triplets comprising anchor, positive, and negative samples to train our model effectively. This approach encourages the network to learn embeddings where genuine and manipulated images of the same individual are mapped to distinct regions in the feature space. We formed the triplets as follows:

- *Anchor*: An image (real or fake) of a specific individual.
- *Positive*: Same individual, same authenticity as anchor, different timestamp.
- *Negative*: Same individual, same timestamp as anchor, opposite authenticity.

We partitioned the real and fake samples into two halves for each identity. Denote these partitions as  $R_1$ ,  $R_2$  (real samples) and  $F_1$ ,  $F_2$  (fake samples). We ensured that the anchor and negative samples were temporally aligned. We then systematically generated all possible triplet configurations for each identity, alternating which partition serves as anchor, positive, and negative samples, as illustrated in Algorithm 1.

---

#### Algorithm 1 Triplet Generation

---

```

function FORMTRIPLETS(RealSamples, FakeSamples)
  Divide RealSamples into two parts:  $R_1$  and  $R_2$ 
  Divide FakeSamples into two parts:  $F_1$  and  $F_2$ 
  Initialize empty collections for Triplets and Labels
  for each identity do
    Add  $(R_1, R_2, F_1)$  to Triplets with label  $(0, 0, 1)$ 
    Add  $(F_1, F_2, R_1)$  to Triplets with label  $(1, 1, 0)$ 
    Add  $(R_2, R_1, F_2)$  to Triplets with label  $(0, 0, 1)$ 
    Add  $(F_2, F_1, R_2)$  to Triplets with label  $(1, 1, 0)$ 
  end for
  return Triplets, Labels
end function

```

---

This triplet formation constitutes the main novelty of our approach, allowing us to control for identity, scene, and temporal factors. We aim to disentangle manipulation cues from identity or scene-specific information by maintaining consistent identity and temporal alignment while varying authenticity.

## IV. EXPERIMENTS

We evaluated the proposed method across three experimental scenarios using the FaceForensics++ (FF++) dataset [10] and conducted cross-dataset evaluation on Google's Deepfake Detection (DFD) dataset [66]. This section details the experimental setups, implementation specifics, and the configurations employed for the ablation studies.

#### A. Area Under the Curve

To evaluate the performance of our method, we employ the *Area Under the Receiver Operating Characteristic Curve (AUC)*, a widely used metric in deepfake detection literature. The AUC provides a threshold-independent measure of a classifier's ability to distinguish between real and fake samples.

1) *AUC and the ROC Curve*: The AUC is computed as the integral of the Receiver Operating Characteristic (ROC) curve by computing the area under the ROC curve, which plots the *True Positive Rate (TPR)* against the *False Positive Rate (FPR)* at various classification thresholds.

2) *Computation of AUC*: The ROC curve is integrated using the trapezoidal rule to compute AUC numerically. Given a set of  $n$  thresholds, the AUC is computed using the trapezoidal rule:

$$\text{AUC} = \sum_{k=1}^{n-1} \frac{(FPR_{k+1} - FPR_k) \cdot (TPR_k + TPR_{k+1})}{2} \quad (7)$$

where  $FPR_k$  and  $TPR_k$  represent the values of the False Positive Rate and True Positive Rate at the  $k^{\text{th}}$  threshold.

TABLE II: Breakdown of Benchmarks

Dataset	Pristine	Manipulated	Total
Celeb-DF (v1) [71]	408	795	1,203
Celeb-DF (v2) [71]	590	5,639	6,229
FaceForensics++ [10]	1,000	4,000	5,000
DFD [66]	363	3,000	3,363
DFDC [72]	19,154	86,654	105,808
DFDCP [73]	1,131	4,119	5,250
UADFV [19]	49	49	98

AUC is particularly suitable for deepfake detection due to its robustness to class imbalance and threshold independence.

### B. Datasets

To evaluate the generalization capability of our approach, we employed several benchmark datasets with diverse forgery techniques and quality levels:

*FaceForensics++ (FF++)* [10] contains 1,000 authentic videos alongside their corresponding manipulated versions generated through four distinct facial manipulation techniques: Deepfakes [67], Face2Face [68], FaceSwap [69], and NeuralTextures [70]. The dataset provides videos at multiple compression levels, enabling comprehensive evaluation across diverse quality degradation scenarios.

*Celebrities DeepFake (Celeb-DF)* [71] includes two versions: *Celeb-DF (v1)* consists of 408 real videos and 795 corresponding DeepFake videos, and *Celeb-DF (v2)* significantly expands upon this, comprising 590 real videos and 5,639 corresponding DeepFake videos with minimal visual artifacts. Version 2 is particularly challenging for detection due to its realistic visual quality, closely matching real-world DeepFake videos circulated online.

*DeepFake Detection (DFD)* [66] released by Google contains thousands of Deepfake videos featuring consenting actors in controlled environments. It encompasses various facial expressions, head poses, and lighting conditions.

*DeepFake Detection Challenge (DFDC)* [72] released by Facebook contains over 100,000 videos created using various deepfake techniques. It features diverse subjects across different ages, ethnicities, and lighting conditions. We also use its preview version *DeepFake Detection Challenge Preview (DFDCP)* [73], which contains a smaller subset of videos but maintains similar diversity, making both datasets particularly challenging for generalization.

*University at Albany DeepFake Videos (UADFV)* [19] includes authentic and deepfake manipulated videos specifically curated to expose inconsistencies in head poses—a common artifact in early deepfake synthesis methods.

Table II provides a concise summary of these datasets, including the number of pristine (original) and manipulated (fake) samples.

### C. Evaluation Scenarios

We designed three experimental scenarios to evaluate the effectiveness of the proposed method and analyze the contributions of its components:

- *Scenario 1: Forgery-Specific Training.* Models were trained on only two forgery types (*Deepfakes* and *NeuralTextures*) from the FF++ dataset. The trained models were then tested on all forgery types in FF++ to assess their ability to generalize to unseen manipulation techniques. This scenario examined the impact of the Gradient Reversal Layer (GRL) and Triplet Learning.
- *Scenario 2: Full Dataset Training.* Models were trained and tested on all forgery types in the FF++ dataset. This scenario measured the overall performance of the proposed method when exposed to a diverse set of forgeries during training.
- *Scenario 3: Cross-Dataset Evaluation.* Models trained on the FF++ (HQ) dataset were evaluated on the DFD dataset. This scenario tested the models’ generalization to a completely different dataset with variations in data distribution and manipulation techniques.

### D. Implementation Details

1) *Network Architecture:* We employed both EfficientNet-B4 [9] (CNN) and MARLIN [74] (ViT) as backbone architectures for feature extraction to explore complementary representation capabilities.

We also designed a *NetworkTree* structure to implement the proposed architecture, enabling flexible configuration and easy insertion or modification of components. The overall network architecture forms a tree with the root being the embedding model and two child modules:

- *Forgery Detector:* A binary classifier that predicts whether an input image is real or fake, utilizing the embedding from the feature extractor.
- *Forgery Discriminator:* An auxiliary classifier that predicts the forgery type, used with the GRL to encourage the embedding model to learn features invariant to specific forgery techniques.

TABLE III: Hyperparameter Comparison

Backbone	EfficientNet-B4 [9]	CLIP ViT-L/14 [65]
Learning rate	0.0001	0.00002
Batch size	4	8
Triplet Loss margin ( $m$ )	1.0	1.0
GRL lambda ( $\lambda$ )	1.0	1.0
Forgery loss weight	1.0	0.5
Number of epochs	30	7
Optimizer	Adam [75]	Adam [75]

2) *Training Configuration:* We employed different hyperparameter configurations for our CNN and ViT variants, as shown in Table III. The ViT variant requires a lower learning rate (0.00002 vs. 0.0001) and fewer epochs (7 vs. 30) than the CNN variant due to parameter-efficient fine-tuning. Both variants share identical triplet margin and GRL parameters.

## V. RESULTS

This section presents the experimental results of our framework. The performance is evaluated quantitatively through ablation studies combined with intra-dataset and cross-dataset evaluations. Qualitative analysis is provided through t-SNE visualization of the learned feature embeddings.

TABLE IV: Frame-level AUC ( $\uparrow$ ) Scores of Face Forgery Detectors. The Best Results are Indicated in Bold.

Type	Detector	Backbone	Venue	Intra-Dataset Evaluation on FF++ (HQ)					Cross-Dataset Evaluation					
				FF++ (HQ)	FF-DF	FF-F2F	FF-FS	FF-NT	Celeb-v1	Celeb-v2	DFD	DFDCP	DFDC	UADFV
Naive	MesoNet [76]	Designed CNN	WIFS-2018	0.6077	0.6771	0.6170	0.5946	0.5701	0.7358	0.6091	0.5481	0.5994	0.5560	0.7150
	MesoInception [76]	Designed CNN	WIFS-2018	0.7583	0.8542	0.8087	0.7421	0.6517	0.7366	0.6966	0.6069	0.7561	0.6226	0.9049
	CNN-Aug [3]	ResNet [77]	CVPR-2020	0.8493	0.9048	0.8788	0.9026	0.7313	0.7420	0.7027	0.6464	0.6170	0.6361	0.8739
	Xception [10]	Xception [8]	ICCV-2019	0.9637	0.9799	0.9785	0.9833	0.9385	0.7794	0.7365	0.8163	0.7374	0.7077	0.9379
	EfficientNet-B4 [9]	EfficientNet [9]	ICML-2019	0.9567	0.9757	0.9758	0.9797	0.9308	0.7909	0.7487	0.8148	0.7283	0.6955	0.9472
Frequency	F3Net [29]	Xception [8]	ECCV-2020	0.9635	0.9793	0.9796	0.9844	0.9354	0.7769	0.7352	0.7975	0.7354	0.7021	0.9347
	SPSL [31]	Xception [8]	CVPR-2021	0.9610	0.9781	0.9754	0.9829	0.9299	0.8150	0.7650	0.8122	0.7408	0.7040	0.9424
	SRM [78]	Xception [8]	CVPR-2021	0.9576	0.9733	0.9696	0.9740	0.9295	0.7926	0.7552	0.8120	0.7408	0.6995	0.9427
Spatial	Capsule [79]	CapsuleNet [80]	ICASSP-2019	0.8421	0.8669	0.8634	0.8734	0.7804	0.7909	0.7472	0.6841	0.6568	0.6465	0.9078
	DSP-FWA [81]	Xception [8]	CVPRW-2019	0.8765	0.9210	0.9000	0.8843	0.8120	0.7897	0.6680	0.7403	0.6375	0.6132	0.8539
	Face X-ray [11]	HRNet [82]	CVPR-2020	0.9592	0.9794	<b>0.9872</b>	0.9871	0.9290	0.7093	0.6786	0.7655	0.6942	0.6326	0.8989
	FFD [83]	Xception [8]	CVPR-2020	0.9624	0.9803	0.9784	0.9853	0.9306	0.7840	0.7435	0.8024	0.7426	0.7029	0.9450
	CORE [45]	Xception [8]	CVPRW-2022	0.9638	0.9787	0.9803	0.9823	0.9339	0.7798	0.7428	0.8018	0.7341	0.7049	0.9412
	RECCE [51]	Custom	CVPR-2022	0.9621	0.9797	0.9779	0.9785	0.9357	0.7677	0.7319	0.8119	0.7419	0.7133	0.9446
	UCF [48]	Xception [8]	ICCV-2023	0.9705	0.9883	0.9840	<b>0.9896</b>	0.9441	0.7793	0.7527	0.8074	0.7594	0.7191	0.9528
	LSDA [84]*	EfficientNet [9]	CVPR-2024	0.9549	0.9738	0.9772	0.9691	0.8994	0.8689	0.8014	0.8296	0.7810	0.7330	0.9009
	Effort [85]*	CLIP ViT [65]	ICML-2025	0.9083	<b>0.9907</b>	0.9124	0.9843	0.7459	0.9041	0.8497	0.8991	0.8133	0.8211	0.9659
	Trident (CNN)	EfficientNet [9]	Ours	<b>0.9793</b>	0.9779	0.9843	0.9831	<b>0.9590</b>	0.7609	0.7543	0.8740	<b>0.8403</b>	0.6976	0.9450
	Trident (ViT)	CLIP ViT [65]	Ours	0.9266	0.9899	0.9441	0.9803	0.7920	<b>0.9115</b>	<b>0.8606</b>	<b>0.9204</b>	0.8263	<b>0.8476</b>	<b>0.9666</b>

\*Reproduced with the official implementations due to unavailability of weights on DeepfakeBench.

### A. Quantitative Results

Table IV presents face forgery detection performance under standardized DeepfakeBench [7] evaluations. All the methods are trained on FF++ (HQ) dataset. Cross-dataset results reveal significant performance degradation across methods, highlighting generalization challenges. To ensure a fair comparison, all metrics are obtained through DeepfakeBench’s controlled evaluation setting.

1) *Ablation Studies*: Ablation studies were conducted to assess the impact of key components in the *Trident* framework with the CNN backbone, specifically Triplet Learning (TL), the Gradient Reversal Layer (GRL), the Adversarial Loss (Adv), and the Detached Classification Head (DH). Experiments were performed under two settings:

- 1) training and testing on all FF++ (HQ) dataset forgery types and
- 2) training on only Deepfakes and NeuralTextures manipulations, then testing on all forgery types in the FF++ (HQ) dataset.

a) *Ablation Study on All Forgery Types*: Table V summarizes the model’s training and testing results on all forgery types in the FF++ (HQ) dataset. The baseline model (B) employs the EfficientNet-B4 backbone with a binary classification head. The variants include the addition of Triplet Learning (TL) and the incorporation of the Gradient Reversal Layer (TL+GRL).

TABLE V: Ablation Study on FF++ (HQ) Dataset (All Forgery Types)

Method	AUC ( $\uparrow$ )	LogLoss ( $\downarrow$ )
Baseline	0.9497	0.2915
TL	0.9613	0.2611
TL+GRL	0.9669	0.2946
TL+GRL+DH	<b>0.9793</b>	<b>0.2456</b>

b) *Ablation Study on Deepfakes and NeuralTextures*:

In this setting, the models are trained only on Deepfakes

and NeuralTextures manipulations and tested on all forgery types in the FF++ (HQ) dataset. Table VI presents the results. The methods include the baseline (B), the addition of Triplet Learning (TL), the incorporation of the Gradient Reversal Layer (TL+GRL), the use of Adversarial Loss (TL+Adv), and the inclusion of the Detached Classification Head (TL+GRL+DH).

TABLE VI: Ablation Study on FF++ (HQ) Dataset (Deepfakes and NeuralTextures)

Method	AUC ( $\uparrow$ )	LogLoss ( $\downarrow$ )
Baseline	0.7363	0.9512
TL	0.7506	0.8645
TL+Adv	0.9595	0.3270
TL+GRL	<b>0.9652</b>	0.2662
TL+GRL+DH	0.9646	<b>0.2454</b>

c) *Cross-Dataset Ablation on DFD (HQ)*: To further assess the impact of our proposed components on generalization ability, we conducted an ablation study on the DFD (HQ) dataset with models trained on FF++ (HQ). Table VII presents the results. The *LogLoss* refers to the Binary Cross-Entropy (BCE) loss used for the classification task in both tables.

TABLE VII: Cross-Dataset Ablation on DFD (HQ) Dataset (Training on FF++ (HQ))

Method	AUC ( $\uparrow$ )	LogLoss ( $\downarrow$ )
Baseline	0.7817	0.6897
TL	0.8438	<b>0.5522</b>
TL+GRL	0.8668	0.6076
TL+GRL+DH	<b>0.8740</b>	0.5568

### B. Qualitative Results

To qualitatively assess the discriminative capability of the learned embeddings, we employed t-distributed Stochastic Neighbor Embedding (t-SNE) [86] to project high-dimensional features into two-dimensional space. We present two sets

of visualizations: baseline models in Figure 2 and Triplet Learning-based models in Figure 3. Real and fake samples are represented in blue and orange, respectively. All visualizations are produced from the FF++ (HQ) validation split.

Figure 2 illustrates embeddings for models *without* Triplet Learning (TL). In Fig. 2(a) (Baseline, B), real (blue) and fake (orange) samples overlap considerably, indicating weak discriminative power. Adding Gradient Reversal Layer (GRL) and Detached Head (DH) components (B+GRL+DH, Fig. 2(b)) strengthens real/fake separation but still does not yield five clear clusters for the FF++ forgeries.

In contrast, Fig. 3 shows *TL-based* models. As seen in Fig. 3(a) (TL), incorporating TL creates distinct clusters for each forgery type, though some overlap remains. Employing GRL and DH components alongside TL (TL+GRL+DH, Fig. 3(b)) further separates real from fake while forming five distinct clusters corresponding to the five forgery types present in the FF++ dataset.

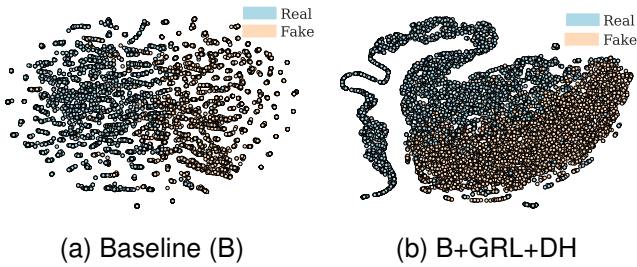


Fig. 2: t-SNE embeddings for models without TL. GRL+DH strengthens real/fake separation but does not yield five distinct forgery clusters.

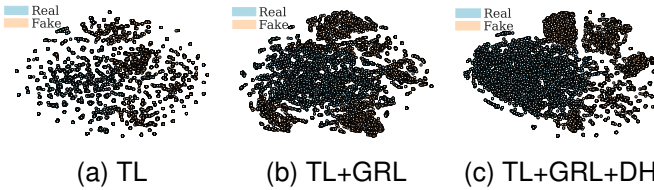


Fig. 3: t-SNE embeddings for TL-based models. TL promotes forgery-type clustering; GRL and DH enhance real/fake separation.

## VI. DISCUSSION

The experimental results presented in Section V provide several important insights into face forgery detection:

a) *Performance Comparison:* The evaluation results in Table IV demonstrate that *Trident* (CNN) achieves 0.9793 AUC on FF++ (HQ), surpassing the previous SOTA approach, UCF (0.9705). In cross-dataset evaluation, our CNN variant achieves notable improvements on DFD (0.8740 vs. Xception’s 0.8163) and DFDCP (0.8403 vs. UCF’s 0.7594). The ViT variant outperforms previous methods on multiple benchmarks, demonstrating the best performance on Celeb-v1 (0.9115 vs. Effort’s 0.9041), Celeb-v2 (0.8606 vs. Effort’s 0.8497), DFD (0.9204 vs. Effort’s 0.8991), DFDC (0.8476 vs.

Effort’s 0.8211), and UADFV (0.9666 vs. Effort’s 0.9659). This consistent performance improvement across different datasets highlights the mitigation effect of our framework against the generalization challenge in face forgery detection.

b) *Component Contributions:* The ablation studies quantify each component’s impact. From Tables V and VI, Triplet Learning provides a foundational improvement, raising the baseline AUC from 0.9497 to 0.9613 with all forgery types and from 0.7363 to 0.7506 with limited forgery types. The most significant gain occurs when combining TL with adversarial techniques (GRL/Adv): when trained only on Deepfakes and NeuralTextures but tested on all forgery types, TL+GRL achieves 0.9652 AUC—a 22.89% increase over the baseline’s 0.7363. This demonstrates the framework’s ability to learn forgery-agnostic features. The cross-dataset results in Table VII reinforce this finding, showing that TL+GRL+DH achieves 0.8740 AUC on DFD, surpassing the baseline by 11.8%.

c) *Feature Space Organization:* The t-SNE visualizations in Figures 2 and 3 reveal the models’ discriminative capabilities. Models without Triplet Learning show considerable overlap between real and fake samples, even with GRL and DH components. TL-based models produce more distinct clustering by forgery type, with the full *Trident* configuration (TL+GRL+DH) achieving both clear real/fake separation and distinct clustering for different forgery techniques.

d) *CNN vs. ViT Backbone:* The results reveal distinct performance patterns between backbone architectures. The CNN-based *Trident* achieves superior intra-dataset performance while demonstrating strong results on DFD and DFDCP benchmarks. In contrast, the ViT-based variant excels across most cross-dataset scenarios, showing consistent improvement on the majority of evaluation benchmarks. This performance divergence suggests that each architecture captures different aspects of forgery artifacts. The CNN’s hierarchical feature extraction appears effective for controlled intra-dataset scenarios, while the ViT architecture’s self-attention mechanism and BitFit fine-tuning strategy demonstrate superior transferability across diverse cross-domain scenarios, likely due to better preservation of pretrained representations.

e) *Impact of BitFit Fine-tuning:* The ViT variant’s strong cross-dataset performance highlights the effectiveness of BitFit fine-tuning in preserving learned representations while adapting to forgery detection. By updating only bias parameters, BitFit prevents catastrophic forgetting of abstract visual features learned during CLIP pretraining, enabling better generalization to unseen datasets and forgery types. This parameter-efficient approach proves particularly valuable in scenarios with limited training data or when adapting to new domains.

f) *Generalization Challenges:* Despite significant improvements achieved, cross-dataset performance reveals ongoing generalization challenges in deepfake detection. The performance variation across different test sets suggests that current methods, including ours, remain sensitive to dataset-specific characteristics such as compression artifacts, capture conditions, and forgery implementation details. However, our *Trident* framework demonstrates more consistent cross-domain performance compared to existing approaches, indicating that



controlled triplet learning combined with domain-adversarial training provides a promising direction for addressing these fundamental challenges in forgery detection generalization.

## VII. CONCLUSION

We introduce *Trident*, a forgery detection framework that integrates triplet learning and domain-adversarial training to address the growing sophistication and diversity of face forgery methods. Our approach isolates subtle forgery-specific features by leveraging triplet learning with carefully curated samples that share identity and scene but differ in authenticity. This controlled representation learning strategy produces robust embeddings that retain discriminative power across forgery techniques. Simultaneously, our domain-adversarial training scheme with a dedicated forgery discriminator produces forgery-agnostic embeddings that capture generalizable manipulation markers.

We implement two backbone variants: EfficientNet and CLIP ViT with BitFit fine-tuning. The parameter-efficient BitFit approach prevents catastrophic forgetting while enabling effective adaptation to forgery detection tasks with minimal computational overhead.

Experiments demonstrate that CNN-based *Trident* achieves state-of-the-art FF++ (HQ) performance, surpassing existing methods on DeepfakeBench. Our ViT variant demonstrates superior cross-dataset generalization, outperforming existing methods on multiple benchmarks, including Celeb-v1, Celeb-v2, DFD, DFDC, and UADFV. The ablation studies confirm that combining triplet learning, gradient reversal layers, and classifier head detachment significantly enhances detection performance. The t-SNE visualizations further validate our approach by demonstrating a clear separation between real and fake samples with distinct clustering by forgery type.

Our results reveal complementary strengths between backbone architectures: CNN-based *Trident* excels in intra-dataset scenarios with strong performance on specific benchmarks, while ViT-based *Trident* demonstrates superior transferability across diverse cross-domain scenarios. The effectiveness of BitFit fine-tuning in preserving pretrained representations while adapting to forgery detection highlights the value of parameter-efficient approaches in this domain.

While *Trident* shows notable improvements in cross-dataset performance, the remaining gap between intra-dataset and cross-dataset results indicates persistent generalization challenges in face forgery detection. Future research directions include incorporating temporal cues, multi-modal inputs, and advanced metric learning approaches. Investigating adaptive triplet sampling strategies or dynamic margin selection could refine the embedding space and address the remaining generalization challenges.

## REFERENCES

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, ser. NIPS '14, 2014, pp. 2672–2680. **1**
- [2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, ser. NIPS '20, vol. 33, 2020, pp. 6840–6851. **1**
- [3] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '20, 2020, pp. 8695–8704. **1, 7**
- [4] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '15, 2015, pp. 815–823. **1, 3**
- [5] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1180–1189. [Online]. Available: <https://proceedings.mlr.press/v37/ganin15.html> **1, 2, 4**
- [6] E. Ben Zaken, Y. Goldberg, and S. Ravfogel, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1–9. [Online]. Available: <https://aclanthology.org/2022.acl-short-1/> **1, 4**
- [7] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "DeepfakeBench: A comprehensive benchmark of deepfake detection," in *Advances in Neural Information Processing Systems*, ser. NIPS '23, vol. 36, 2023, pp. 4534–4565. **2, 7**
- [8] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '17, 2017, pp. 1251–1258. **2, 7**
- [9] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning*, ser. ICML '19, 2019, pp. 6105–6114. **2, 4, 6, 7**
- [10] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ser. ICCV '19, 2019, pp. 1–11. **2, 5, 6, 7**
- [11] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '20, 2020, pp. 5001–5010. **2, 7**
- [12] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 103–120. [Online]. Available: [https://doi.org/10.1007/978-3-030-58574-7\\_7](https://doi.org/10.1007/978-3-030-58574-7_7) **2**
- [13] X. Li, R. Ni, P. Yang, Z. Fu, and Y. Zhao, "Artifacts-disentangled adversarial learning for deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1658–1670, 2023. **2**
- [14] D. Nguyen, N. Mejri, I. P. Singh, P. Kuleshova, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada, "LAA-Net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '24, 2024, pp. 17395–17405. **2**
- [15] Z. Yang, J. Liang, Y. Xu, X.-Y. Zhang, and R. He, "Masked relation learning for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1696–1708, 2023. **2**
- [16] D. Dagar and D. K. Vishwakarma, "Shallowfake and deepfake image manipulation localization using noise and RGB-based dual branch method," *Signal, Image and Video Processing*, vol. 18, pp. 7065–7077, 2024. **2**
- [17] A. Yadav and D. K. Vishwakarma, "Aw-msa: Adaptively weighted multi-scale attentional features for deepfake detection," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107443, 01 2024. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0952197623016275> **2**
- [18] M. Long, Z. Liu, L.-B. Zhang, and F. Peng, "Lgdf-net: Local and global feature-based dual-branch fusion networks for deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 6, pp. 5489–5500, 2025. **2**
- [19] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP '19, 2019, pp. 8261–8265. **2, 6**
- [20] A. Haliassos, K. Vougioukas, S. Petridis, and M. 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*, ser. CVPR '21, 2021, pp. 5039–5049. **2**

- [21] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, and L. Ma, "Spatiotemporal inconsistency learning for deepfake video detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '21, 2021, pp. 10 343–10 352. [2](#)
- [22] I. Amerini, R. Caldelli, and F. Picchioni, "Deepfake video detection through optical flow based CNN," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, ser. ICCVW '19, 2019, pp. 1205–1207. [2](#)
- [23] I. Choi, Y. Kim, and S. Hwang, "Exploiting inconsistencies in stylegan latent space for deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '24, 2024. [2](#)
- [24] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, and L. Nie, "Voice–Face homogeneity tells deepfake," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 3, 2024. [2](#)
- [25] Y. Yu, X. Zhao, R. Ni, S. Yang, Y. Zhao, and A. C. Kot, "Augmented multi-scale spatiotemporal inconsistency magnifier for generalized deepfake detection," *IEEE Transactions on Multimedia*, vol. 25, pp. 8487–8498, 2023. [2](#)
- [26] W. Lu, L. Liu, B. Zhang, J. Luo, X. Zhao, Y. Zhou, and J. Huang, "Detection of deepfake videos using long-distance attention," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, pp. 9366–9379, 2024. [2](#)
- [27] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, and R. He, "Tall: Thumbnail layout for deepfake video detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 22 658–22 668. [2](#)
- [28] M. Li, B. Liu, Y. Hu, L. Zhang, and S. Wang, "Deepfake detection using robust spatial and temporal features from facial landmarks," in *Proc. Int. Workshop on Biometrics and Forensics (IWBF)*, 2021, pp. 1–6. [2](#)
- [29] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 86–103. [Online]. Available: [https://doi.org/10.1007/978-3-030-58610-2\\_6](https://doi.org/10.1007/978-3-030-58610-2_6) [2](#), [7](#)
- [30] J. Li, H. Xie, J. Li, Z. Wang, and Y. 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*, ser. CVPR '21, 2021, pp. 6458–6467. [2](#)
- [31] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. 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*, ser. CVPR '21, 2021, pp. 772–781. [2](#), [7](#)
- [32] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML '20. JMLR.org, 2020, pp. 3247–3258, article no. 304. [2](#)
- [33] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '24, June 2024, pp. 28 130–28 139. [2](#)
- [34] —, "Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, ser. AAAI '24, vol. 38, no. 5, 2024, pp. 5052–5060. [2](#)
- [35] Z. Wang, Z. Cheng, J. Xiong, X. Xu, T. Li, B. Veeravalli, and X. Yang, "A timely survey on vision transformer for deepfake detection," 2024. [Online]. Available: <https://arxiv.org/abs/2405.08463> [2](#)
- [36] C. Zhao *et al.*, "ISTVT: Interpretable spatial-temporal video transformer for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1335–1348, 2023. [2](#)
- [37] J. Wang, Z. Wu, J. Chen, and Y. Jiang, "M2TR: Multi-modal multi-scale transformers for deepfake detection," *CoRR*, vol. abs/2104.09770, 2021. [Online]. Available: <https://arxiv.org/abs/2104.09770> [2](#)
- [38] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Protecting celebrities from deepfake with identity consistency transformer," 2022. [Online]. Available: <https://arxiv.org/abs/2203.01318> [2](#)
- [39] W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, and N. Yu, "UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," 2022. [Online]. Available: <https://arxiv.org/abs/2210.12752> [2](#)
- [40] S. A. Khan and H. Dai, "Video transformer for deepfake detection with incremental learning," *CoRR*, vol. abs/2108.05307, 2021. [Online]. Available: <https://arxiv.org/abs/2108.05307> [2](#)
- [41] D. Wodajo, S. Atnafu, and Z. Akhtar, "Deepfake video detection using generative convolutional vision transformer," 2023. [Online]. Available: <https://arxiv.org/abs/2307.07036> [2](#)
- [42] A. Luo, R. Cai, C. Kong, Y. Ju, X. Kang, J. Huang, and A. C. Kot, "Forgery-aware adaptive learning with vision transformer for generalized face forgery detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 5, pp. 4116–4129, 2025. [2](#)
- [43] D. Dagar and D. K. Vishwakarma, "Tex-ViT: A generalizable, robust, texture-based dual-branch cross-attention deepfake detector," *arXiv preprint arXiv:2408.16892*, 2024. [2](#)
- [44] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to GANs: learning and analyzing gan fingerprints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ser. ICCV '19, 2019, pp. 7556–7566. [2](#)
- [45] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao, "CORE: Consistent representation learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, ser. CVPRW '22, 2022, pp. 12–21. [2](#), [7](#)
- [46] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ser. ICCV '21, 2021, pp. 15 023–15 033. [2](#)
- [47] X. Huang, F. Xue, B. Fan, L. Zhong, Y. Fu, and Q. Tian, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '23, 2023, pp. 3994–4004. [2](#)
- [48] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: Uncovering common features for generalizable deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ser. ICCV '23, 2023, pp. 22 355–22 366. [2](#), [7](#)
- [49] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '23, 2023, pp. 24 480–24 489. [2](#)
- [50] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '22, 2022, pp. 18 720–18 729. [2](#)
- [51] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '22, 2022, pp. 4113–4122. [2](#), [7](#)
- [52] K. Zhang, Z. Hou, Z. Hua, Y. Zheng, and L. Y. Zhang, "Boosting deepfake detection generalizability via expansive learning and confidence judgement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 1, pp. 953–966, 2025. [2](#)
- [53] Y. Yu, X. Liu, R. Ni, S. Yang, Y. Zhao, and A. C. Kot, "PVASS-MDD: Predictive visual–audio alignment self-supervision for multimodal deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 17–30, 2024. [2](#)
- [54] X. Liu, Y. Yu, X. Li, and Y. Zhao, "MCL: Multimodal contrastive learning for deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2803–2813, 2024. [2](#)
- [55] S. Fung, X. Lu, C. Zhang, and C. Li, "DeepfakeUCL: Deepfake detection via unsupervised contrastive learning," *CoRR*, vol. abs/2104.11507, 2021. [Online]. Available: <https://arxiv.org/abs/2104.11507> [2](#)
- [56] Y. Xu, K. Raja, and M. Pedersen, "Supervised contrastive learning for generalizable and explainable deepfakes detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, ser. WACVW '22, 2022, pp. 379–389. [2](#)
- [57] Z. Gu, T. Yao, Y. Chen, S. Ding, and L. Ma, "Hierarchical contrastive inconsistency learning for deepfake video detection," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*. Berlin, Heidelberg: Springer-Verlag, 2022, pp. 596–613. [Online]. Available: [https://doi.org/10.1007/978-3-031-19775-8\\_35](https://doi.org/10.1007/978-3-031-19775-8_35) [2](#)
- [58] B. Liang, Z. Wang, B. Huang, Q. Zou, Q. Wang, and J. Liang, "Depth map guided triplet network for deepfake face detection," *Neural Networks*, vol. 159, pp. 34–42, 2023. [2](#)
- [59] A. Kumar, A. Bhavsar, and R. Verma, "Detecting deepfakes with metric learning," in *Proceedings of the 8th International Workshop on Biometrics and Forensics*, ser. IWBF '20, 2020, pp. 1–6. [2](#)

- [60] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18, 2018, pp. 849–857. [2](#)
- [61] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, Jun. 2009. [3](#)
- [62] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, ser. CVPR '05, vol. 1. IEEE, 2005, pp. 539–546. [3](#)
- [63] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '16, 2016, pp. 4004–4012. [3](#)
- [64] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *Advances in Neural Information Processing Systems*, ser. NIPS '93, vol. 6, 1993. [3](#)
- [65] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, ser. ICML '21, 2021, pp. 8748–8763. [4](#), [6](#), [7](#)
- [66] Google and Jigsaw, "DeepFake detection dataset," 2019, <https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html>, Accessed: 2025-02-16. [5](#), [6](#)
- [67] Deepfakes Community, "Deepfakes," <https://github.com/deepfakes/faceswap>, 2018, accessed: 2018-10-29. [6](#)
- [68] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '16, 2016, pp. 2387–2395. [6](#)
- [69] M. Kowalski, "FaceSwap," <https://github.com/MarekKowalski/FaceSwap/>, 2018, accessed: 2018-10-29. [6](#)
- [70] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics*, vol. 38, no. 4, 2019. [6](#)
- [71] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '20, 2020, pp. 3207–3216. [6](#)
- [72] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397*, 2020. [6](#)
- [73] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," *arXiv preprint arXiv:1910.08854*, 2019. [6](#)
- [74] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, R. Haffari, and M. Hayat, "Marlin: Masked autoencoder for facial video representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '23, 2023, pp. 1493–1504. [6](#)
- [75] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference for Learning Representations*, ser. ICLR '15, 2015. [Online]. Available: <https://arxiv.org/abs/1412.6980> [6](#)
- [76] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, ser. WIFS '18. IEEE, 2018, pp. 1–7. [7](#)
- [77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '16, 2016, pp. 770–778. [7](#)
- [78] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '21, 2021, pp. 16 317–16 326. [7](#)
- [79] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-Forensics: Using capsule networks to detect forged images and videos," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP '19, 2019, pp. 2307–2311. [7](#)
- [80] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, ser. NIPS '17, vol. 30, 2017, pp. 3859–3869. [7](#)
- [81] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, ser. CVPRW '19, June 2019, pp. 46–52. [7](#)
- [82] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020. [7](#)
- [83] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '20, 2020, pp. 5781–5790. [7](#)
- [84] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu, "Transcending forgery specificity with latent space augmentation for generalizable deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR '24, 2024, pp. 8984–8994. [7](#)
- [85] Z. Yan, J. Wang, P. Jin, K.-Y. Zhang, C. Liu, S. Chen, T. Yao, S. Ding, B. Wu, and L. Yuan, "Orthogonal subspace decomposition for generalizable AI-generated image detection," in *Proceedings of the International Conference on Machine Learning*, ser. ICML '25, 2025. [7](#)
- [86] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov 2008. [7](#)