

Trident: Detecting Face Forgeries With Domain-Adversarial Triplet Learning

Mustafa Hakan Kara , Aysegul Dundar , and Uğur Gündükbay , *Senior Member, IEEE*

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 trained on 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 within a Siamese network architecture to enhance 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 toward forgery-agnostic representations, improving its robustness to unseen manipulations. In addition, we prevent gradient flow from the classifier head to the embedding model, thereby avoiding overfitting induced by artifacts specific to certain forgeries. Comprehensive evaluations across multiple benchmarks and ablation studies demonstrate the effectiveness of our framework.

Impact Statement—Reliable detection of face forgeries is critical for the responsible use of digital media across online platforms, journalism, and identity-sensitive applications. Many existing detection systems perform well under controlled conditions but degrade when confronted with previously unseen manipulation methods, limiting their effectiveness in practice. This work addresses that limitation by enhancing the robustness of face-forgery detection and reducing dependence on incidental cues specific to certain forgery methods. Improved robustness across datasets enables more dependable automated forensic analysis in settings where errors may affect individuals, institutions, and public trust. Beyond measured performance gains, the study provides insight into how training design affects a detector's effectiveness as manipulation techniques evolve. By strengthening generalization rather than tuning for narrow scenarios, this research supports the development of media verification technologies that remain reliable as face forgery methods continue to evolve.

Index Terms—Deepfake detection (DFD), domain-adversarial training, domain generalization, generalizable face forgery detection, triplet learning (TL).

I. INTRODUCTION

ADVANCES in deep generative models, 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, supervised approaches have limitations in

- 1) generalization across diverse manipulation techniques;
- 2) detection of previously unseen forgery methods [3].

A key reason for these limitations is that models often rely on dataset-specific or identity-specific shortcuts rather than manipulation-related cues, which weakens generalization. To address these challenges, we present **triplet learning-based deepfake detection network** (*Trident*), a triplet learning (TL)-based framework 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 *controlled* 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 such as background and identity. This design ensures that the learned embeddings capture the intrinsic differences between genuine and manipulated samples, enhancing the generalization and robustness of the detection process.

Trident also leverages domain-adversarial training [5] with a forgery discriminator, which attempts to classify the forgery category given an embedding. With the gradient reversal layer (GRL), this interaction becomes a minimax optimization problem in which the encoder is driven to suppress forgery-related information. This adversarial interplay prompts the embedding generator to produce representations that are 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. The GRL formulation also avoids the instability often associated with GAN-style adversarial setups.

In addition, we detach the Siamese network embeddings before passing them to the classifier head, ensuring that the

Received 26 March 2026; revised 17 May 2026; accepted 9 June 2026. Date of publication 15 June 2026; date of current version 30 September 2026. This article was recommended for publication by Associate Editor Zahid Akhtar upon evaluation of the reviewers' comments. (Corresponding author: Uğur Gündükbay.)

The authors are with the Department of Computer Engineering, Bilkent University, Ankara 06800, Türkiye (e-mail: hakan.kara@bilkent.edu.tr; adundar@cs.bilkent.edu.tr; gudukbay@cs.bilkent.edu.tr).

Digital Object Identifier 10.1109/TAI.2026.3703428

classifier’s gradients do not backpropagate through the embedding network. This approach maintains the integrity of the embedding space learned through triplet loss, enabling it to focus more on capturing generalizable forgery artifacts rather than classifier-driven shortcuts.

For our transformer-based variant, we employ BitFit [6], a parameter-efficient fine-tuning (PEFT) approach that prevents catastrophic forgetting while enabling efficient adaptation to the forgery detection task via bias-only updates.

Our contributions to face forgery detection are as follows.

- 1) A novel controlled TL formulation that disentangles manipulation cues from identity and scene-specific information by maintaining consistent identity and temporal alignment.
- 2) The introduction of an adversarially trained forgery discriminator, guiding the embeddings toward forgery-agnostic representations that remain effective against unseen manipulations.
- 3) Detaching backbone embeddings from the binary classifier head to further increase generalization.
- 4) 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 TL, domain-adversarial training, and PEFT to detect the evolving landscape of face forgeries, thereby contributing to more robust methods for preserving the authenticity of digital media.

II. RELATED WORK

Face forgery detection has evolved significantly to address increasingly sophisticated manipulation techniques [7]. With the widespread adoption of deep learning and convolutional neural networks, face forgery detection techniques have advanced significantly. We provide a representative sample of these methods. For a detailed coverage, please refer to the surveys [8], [9].

A. Spatial-Domain Methods

While early approaches to face forgery detection relied on detecting specific artifacts from generation algorithms—with Xception [10] and EfficientNet [11] serving as robust baselines [12]—the focus has recently shifted toward blending and synthesis anomalies. For instance, *Face X-Ray* [13] targets blending boundaries, while Chai et al. [14] analyze statistical differences between camera-captured and synthesized images.

Recent spatial-domain detectors integrate explicit artifact reasoning with fine-grained attention. Several methods employ local strategies, such as voting over semantic patches and adaptive aggregation from manipulation-specific experts. Others focus on region-specific anomalies by identifying likely tampered regions and leveraging multiview analysis. To move beyond binary supervision, some approaches couple reconstruction with classification. Furthermore, decomposition strategies have gained traction, separating graphics components to mine

forgery-sensitive details, and related approaches leverage residual cues guided by identity-level discrepancies. Zhang et al. combine fine-grained spatial cues with high-frequency noise inconsistencies through a two-stream design and a dedicated fusion module [15].

B. Temporal and Frequency Domain Methods

Temporal detectors exploit inconsistencies across video frames, such as abnormal head-pose dynamics [16] and incoherent mouth movements [17]. To capture these cues, Gu et al. [18] proposed spatiotemporal inconsistency learning, while Amerini et al. [19] utilized optical-flow convolutional neural networks (CNNs) to uncover compression-robust artifacts. Chhabra et al. address low-quality deepfake detection (DFD) in the compressed domain by enforcing sensitivity to subtle, previously unseen artifacts [20]. Recent advances focus on refined feature decoupling and auxiliary supervision (e.g., biometric cues) to improve robustness, while Choi et al. [21] detect style-latent incoherence between frames. Pintelas et al. explore quantization-aware 3-D CNNs with a gradual-unfreezing schedule, targeting deepfake cues that may become more salient under discrete representations [22].

Complementary to spatial and temporal analysis, frequency-domain methods analyze spectral artifacts introduced during the synthesis process. F³-Net [23] mines frequency-aware clues, Li et al. [24] learn discriminative features across bands, and others exploit phase discrepancies [25] or high-frequency artifacts linked to GAN up-sampling [26].

C. Vision Transformer (ViT)-Based Forgery Detection

ViTs have emerged as powerful architectures for DFD due to their ability to capture long-range dependencies and fine-grained features [27]. Interpretable spatial-temporal video transformer (ISTVT) [28] combines spatial and temporal attention, while Xu et al. [29] introduced thumbnail layout (TALL)-Swin, which transforms video clips into a compact thumbnail layout to efficiently preserve spatiotemporal dependencies. Multi-modal multi-scale transformer (M2TR) [30] analyzes images at multiple scales through parallel transformer branches. Identity consistency transformer (ICT) [31] verifies identity consistency, and unsupervised inconsistency-aware vision transformer (UIA-ViT) [32] models feature distributions without pixel-level masks. Khan et al. [33] fuse parametric texture maps with transformers for pose-invariant detection, and generative convolutional vision transformer (GenConViT) [34] combines generative priors. Related hybrid designs also pair CNN feature maps with transformer self-attention to capture both local artifacts and broader contextual dependencies in videos [35].

D. Generalization-Focused Methods

Achieving robustness across unseen manipulations remains a critical challenge. Early generalization works focused on learning GAN fingerprints [36] or enforcing representation consistency [37], [38]. Closely related, Lai et al. [39] fuse content features with sampled style representations to retain

task-relevant domain cues while suppressing background and identity confounders. To mitigate preprocessing dependencies, some approaches curate training data by favoring samples that yield stable predictions across normalization changes. Architecture design choices and search-based strategies have also been explored to improve cross-dataset robustness.

Several methods enhance robustness by synthesizing training samples or reconstructing real faces. For instance, some approaches cast cross-dataset detection as an anomaly detection task, using reconstruction to identify unseen fakes via latent-space inconsistencies. Similarly, self-blended images (SBIs) [40] employ self-blending, while reconstruction-classification learning (RECCE) [41] reconstructs authentic faces to highlight anomalies. Other strategies include disentangling identity information [42], decomposing images into forgery-relevant partitions [43], or employing frozen encoders with classical classifiers [44]. Mixture-of-experts designs over a contrastive language-image pre-training (CLIP) backbone disentangle domain-specific from domain-invariant cues, coupling per-domain expert layers with meta-learned updates to improve cross-domain transfer [45]. In addition, some methods address demographic attribute bias to encourage more consistent performance across diverse groups, including kernel-based conditional distribution alignment to improve group fairness [46].

E. Contrastive and Metric Learning Approaches

Contrastive learning enhances generalizability by enforcing tighter intra-class compactness and inter-class separation. Approaches range from unsupervised schemes, such as deepfake detection via unsupervised contrastive learning (DeepfakeUCL) [47], to supervised contrastive learning [48] and hierarchical video contrast [49]. Metric learning variants, such as depth-guided triplets [50] and triplet networks for high compression [51], further refine the feature space. Wu et al. introduce language-guided contrastive supervision for synthetic image detection, improving transfer to unseen generators [52]. Prototype-based formulations have also been used to target cross-domain robustness, including multimodal prototype matching and contrastive prototype learning for deepfake videos [53], [54].

F. Domain-Adversarial Training

Domain-adversarial training aims to learn domain-invariant features by confusing a domain discriminator. Originating from the GRL [5], this concept was adapted for fake news detection in event adversarial neural network (EANN) [55]. Our work builds on these principles, employing adversarial and triplet objectives to realize forgery-agnostic representations.

G. Vision-Language and LLM-Agent Approaches

A recent line of work uses large language models and vision-language backbones for forensic reasoning. Lai et al. [56] instantiate large language model (LLM)-driven agents that simulate the forgery-creation process within a social-media environment, producing training samples annotated for text and image consistency. From a complementary angle, Lai et al. [57]

examine how the number of visual tokens supplied to a forensic vision-language model affects detection quality, finding that additional tokens improve accuracy with diminishing returns beyond a moderate budget.

III. METHOD

Existing DFD methods often struggle to generalize to unseen forgery techniques, as they tend to rely on artifacts that are specific to the manipulation methods observed during training. To address this limitation, we propose a representation-learning framework that encourages the model to focus on manipulation-induced inconsistencies rather than on forgery-specific signatures.

Our approach consists of three key components. First, we introduce a controlled TL formulation for DFD. By constructing triplets that preserve the same person identity and scene, we isolate forgery-induced variations and encourage the model to capture subtle artifacts introduced by the manipulation process rather than incidental changes in identity or background.

Second, we incorporate a domain-adversarial training objective to promote forgery-agnostic representations. A forgery discriminator attempts to classify the manipulation technique from the learned embeddings, while the embedding generator is trained to produce features that are invariant to these technique-specific cues.

Finally, we employ a binary forgery detector that predicts whether a sample is real or fake, while decoupling its training from the representation learning process. Specifically, gradients from the classifier are prevented from propagating back into the embedding generator, ensuring that the embedding space remains governed by the metric and adversarial objectives rather than the binary decision boundary.

The following sections describe these components in detail, including the domain-adversarial training setup and the detached classifier design used for robust forgery detection.

A. TL Framework

Trident employs a TL paradigm [58], [4], extending contrastive learning and Siamese network concepts. In contrastive learning [59], [60], models are trained to pull similar samples closer and push dissimilar ones apart in the feature space, typically via pairwise contrastive loss. Siamese networks [61] feed pairs of samples through identical networks with shared weights to measure similarity, while TL [4] introduces a third element, forming anchor-positive-negative triplets to refine feature discrimination.

We leverage this idea to learn 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. In our framework, we propose a controlled triplet-formation strategy, which constitutes a key methodological contribution that distinguishes our approach from standard triplet-learning methods. Specifically, we form the triplets as follows.

- 1) *Anchor*: A real or fake frame of a specific individual at a given timestamp t .

Algorithm 1: Temporal Triplet Generation With Partition Permutations.

```

1: function FORMTRIPLETS(Frames_Real, Frames_Fake)
2:   Divide Frames_Real into  $R_1, R_2$  ▷ Partitions of real frames
3:   Divide Frames_Fake into  $F_1, F_2$  ▷ Partitions of fake frames
4:   Initialize empty collections for Triplets and Labels
5:   for each identity do
6:     for each timestamp  $t$  in video do
7:       Add triplet  $(R_1, R_2, F_1)$  with label  $(0, 0, 1)$ 
8:       ▷ Anchor= $R_1$ , Positive= $R_2$  (different partition),
9:       ▷ Negative= $F_1$  (same timestamp as anchor)
10:      Add triplet  $(F_1, F_2, R_1)$  with label  $(1, 1, 0)$ 
11:      ▷ Anchor= $F_1$ , Positive= $F_2$ , Negative= $R_1$ 
12:      Add triplet  $(R_2, R_1, F_2)$  with label  $(0, 0, 1)$ 
13:      ▷ Anchor= $R_2$ , Positive= $R_1$ , Negative= $F_2$ 
14:      Add triplet  $(F_2, F_1, R_2)$  with label  $(1, 1, 0)$ 
15:      ▷ Anchor= $F_2$ , Positive= $F_1$ , Negative= $R_2$ 
16:     end for
17:   end for
18:   return Triplets, Labels
19: end function

```

2) *Positive*: A frame of the same individual with the same authenticity as the anchor but from a *different timestamp* $t' \neq t$.

3) *Negative*: A frame of the same individual at the *same timestamp* t as the anchor, but with opposite authenticity.

To systematically generate triplets, we partition real and fake frames for each identity into two halves: R_1, R_2 for real; and F_1, F_2 for fake frames. Temporal alignment is enforced between anchor and negative frames to preserve consistent scene and motion cues. We alternate partitions to serve as anchor, positive, and negative to cover all possible configurations, as illustrated in Algorithm 1.

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.

We validate the effectiveness of this temporal triplet formation through ablation studies (Table V). The proposed *aligned sampling* consistently outperforms other options. By maintaining consistent identity and temporal alignment while varying authenticity, our approach disentangles manipulation cues from identity- or scene-specific information.

B. Domain-Adversarial Training Setup With Gradient Reversal

Our second key contribution is the application of domain-adversarial training to multiclass face forgery detection, extending a technique typically used for two-class source-target problems in the domain adaptation literature. Specifically, our framework employs a domain-adversarial training to learn forgery-agnostic representations that generalize across different manipulation techniques. The core principle involves a min-max game between the embedding generator and a forgery discriminator. The embedding generator aims to produce informative features for distinguishing between real and fake data. However, it is uninformative for forgery classification, whereas

the forgery discriminator aims 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 that produces logits for each forgery category (Deepfakes, Face2Face, FaceSwap, and NeuralTextures). Through this adversarial interaction, the model learns to capture broad real-versus-fake distinctions while avoiding overfitting to particular forgery methods.

The GRL [5] implements this 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 λ 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, λ , 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 multiclass 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. This domain-adversarial training effectively aligns the generator with producing domain-invariant features, preventing overfitting to known forgery types and enabling robustness to unseen manipulations, which, combined with our multiclass extension, constitutes a primary contribution of our approach in face forgery detection.

C. Forgery Detector With Classifier Head Detachment

Our primary objective is robust forgery detection that generalizes to previously unseen manipulation types. To this end, we first learn a representation that captures forgery-related cues while suppressing domain-specific artifacts. Specifically, the encoder-embedder is trained using a combination of triplet loss and domain-adversarial loss as explained in previous subsections, which structures the embedding space according to similarity relationships while encouraging domain-invariant representations. On top of the learned embeddings, we employ a forgery detector, a binary classification head that predicts whether an input sample is real or fake. While the detector provides the final prediction for the task, directly optimizing the encoder with the binary classification objective can cause the representation to overfit to the decision boundary induced by the training forgery types.

As our third contribution, we introduce a decoupled training strategy that explicitly separates the binary forgery-detection objective from the representation-learning process. During

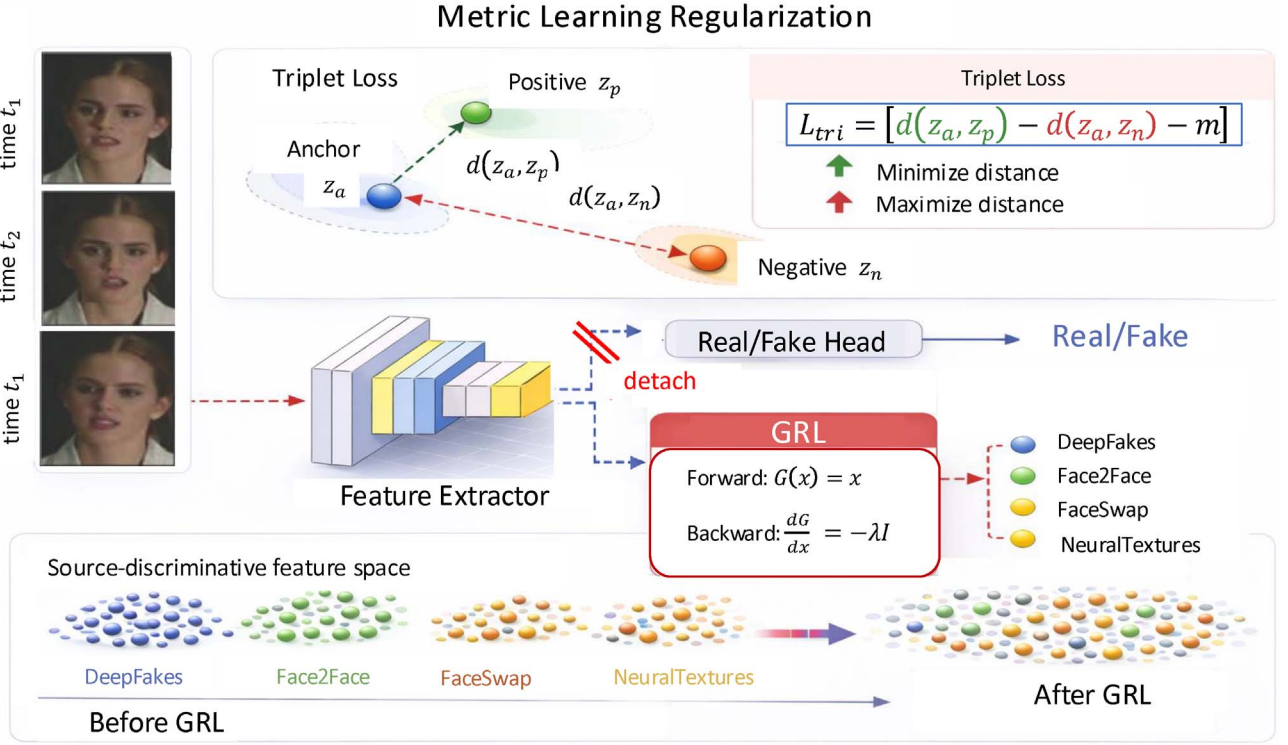


Fig. 1. Overview of our proposed framework for generalized forgery detection. The framework is optimized using three complementary objectives. First, a triplet loss with our controlled sampling strategy encourages the model to capture subtle forgery artifacts by comparing samples that share the same identity and scene while differing in authenticity. Second, a domain-adversarial objective (GRL) with a forgery discriminator promotes domain confusion, encouraging the embeddings to become invariant to specific manipulation techniques. Finally, the embeddings are fed to a binary forgery detector that predicts whether a sample is real or fake. During training, gradients from this classifier are detached from the encoder, ensuring that the embedding space is shaped solely by the metric and adversarial objectives, resulting in more transferable and forgery-agnostic representations.

training, we block the backpropagation from the forgery detector into the shared encoder, preventing the binary classification loss from directly shaping the embedding space. Consequently, the representation is governed solely by the metric and domain-adversarial objectives, which encourage a more structured and domain-agnostic embedding space. This stop-gradient design prevents the encoder from collapsing toward a representation optimized only for binary separability on the training manipulations. Instead, the embeddings organize samples according to forgery-related similarity relationships, resulting in a more stable and transferable representation. Empirically, this decoupling significantly improves the model’s ability to detect previously unseen types of forgery.

D. Model Architecture

Our model architecture, as shown in Fig. 1, features a shared encoder network with a triplet-based learning setup to distinguish real samples from fake ones. We fine-tune two backbone variants for comprehensive evaluation: EfficientNet-B4 [11], a CNN-based architecture commonly preferred across face forgery detection methods, and CLIP ViT-L/14 [62], OpenAI’s transformer-based foundational model pretrained on large-scale visual and textual data.

For the ViT backbone, we employed BitFit [6], a 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

TABLE I
BACKBONE COMPARISON OF EFFICIENTNET-B4 [11] AND CLIP ViT-L/14 [62]

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

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-tune in our experiments.

E. 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 α and β are weighting factors that balance the contributions of the individual loss components.

1) *BCE Loss*: The BCE 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 the distinctiveness of the features.

3) *Forgery Discriminator Loss*: The forgery discriminator loss $\mathcal{L}_{\text{forgery}}$ employs multiclass 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 representations that are forgery-agnostic.

IV. EXPERIMENTS

This section presents the datasets, evaluation metrics, experimental setup, implementation details, and training configurations used in our experiments.

A. Metrics

To evaluate the performance of our method, we employ the area under the receiver operating characteristic curve (AUC), a widely used metric in DFD literature, and the Accuracy of the detectors. AUC provides a threshold-independent measure of a classifier's ability to distinguish between real and fake samples, whereas Accuracy reports classification performance at a fixed decision threshold.

1) *AUC*: 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. The ROC curve is integrated using the trapezoidal

rule to compute AUC numerically. Given a set of n thresholds, the AUC is computed as

$$\text{AUC} = \sum_{k=1}^{n-1} \frac{(\text{FPR}_{k+1} - \text{FPR}_k) \cdot (\text{TPR}_k + \text{TPR}_{k+1})}{2} \quad (7)$$

where FPR_k and TPR_k represent the values of the FPR and TPR at the k^{th} threshold. AUC is particularly suitable for DFD due to its robustness to class imbalance and threshold independence. Additionally, we primarily relied on AUC, as it is the de facto evaluation metric adopted by mainstream DFD benchmarks, which enables direct and fair comparison with prior work.

2) *Accuracy*: We additionally report classification accuracy using a fixed decision threshold of 0.5 on the predicted probability scores. Samples with scores greater than or equal to 0.5 are classified as fake, while those below 0.5 are classified as real, and the final accuracy is computed as the proportion of correctly classified samples.

B. Datasets

To evaluate the generalization capability of our approach, we use several benchmark datasets that feature diverse forgery techniques and varying quality levels.

FaceForensics++ (FF++) [12] contains 1000 authentic videos alongside their corresponding manipulated versions generated through four distinct facial manipulation techniques: Deepfakes [63], Face2Face [64], FaceSwap [65], and Neural-Textures [66]. The dataset provides videos at multiple compression levels, enabling comprehensive evaluation across a range of diverse quality degradation scenarios.

Celebrities DeepFake (Celeb-DF) [67] 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 5639 corresponding DeepFake videos with minimal visual artifacts. Version 2 is particularly challenging to detect due to its realistic visual quality, which closely matches that of real-world DeepFake videos circulating online.

DFD [68] released by Google contains thousands of Deepfake videos featuring consenting actors in controlled environments. It encompasses a range of facial expressions, head poses, and lighting conditions.

DeepFake detection challenge (DFDC) [69] 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 of the *DeepFake detection challenge preview (DFDCP)* [70], which contains a smaller subset of videos but maintains similar diversity, making both datasets particularly challenging for generalization.

University at Albany DeepFake videos (UADFVs) [16] include authentic and deepfake manipulated videos specifically curated to expose inconsistencies in head poses—a common artifact in early deepfake synthesis methods.

WildDeepfake [71] contains 7314 face sequences extracted from 707 deepfake videos collected entirely from the Internet. Unlike datasets generated under controlled conditions, its

TABLE II
BREAKDOWN OF BENCHMARKS

Dataset	Pristine	Manipulated	Total
FaceForensics++ [12]	1000	4000	5000
Celeb-DF (v1) [67]	408	795	1203
Celeb-DF (v2) [67]	590	5639	6229
DFD [68]	363	3000	3363
DFDC [69]	19 154	86 654	105 808
DFDCP [70]	1131	4119	5250
UADFV [16]	49	49	98
WildDeepfake [71]	3805	3509	7314

samples span diverse scenes, identities, compression levels, and synthesis pipelines, providing a stringent in-the-wild evaluation setting.

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

C. Evaluation Scenarios

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

- 1) *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 examines the impact of the GRL and TL.
- 2) *Scenario 2: Full Dataset Training*: Models are trained and tested on all forgery types in the FF++ dataset. This scenario evaluates the overall performance of the proposed method when trained on a diverse set of forgeries.
- 3) *Scenario 3: Cross-Dataset Evaluation*: Models trained on the FF++ (HQ) dataset are evaluated on the DFD dataset. This scenario tests the models' generalization to a completely different dataset with varied data distributions and manipulation techniques.

D. Implementation Details

1) *Network Architecture*: We employed both EfficientNet-B4 [11] (CNN) and CLIP [62] (ViT-L/14) as backbone architectures for feature extraction to explore complementary representation capabilities.

We also design 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, as follows.

- a) *Forgery detector*: A binary classifier that predicts whether an input image is real or fake, utilizing the embedding from the feature extractor.
- b) *Forgery discriminator*: An auxiliary classifier that predicts the forgery type, used with the GRL to encourage the

TABLE III
HYPERPARAMETER COMPARISON OF EFFICIENTNET-B4 [11]
AND CLIP ViT-L/14 [62]

Backbone	EfficientNet-B4	CLIP ViT-L/14
Learning rate	0.0001	0.00002
Batch size	4	8
Triplet Loss margin (m)	1.0	1.0
GRL lambda (λ)	1.0	1.0
Forgery loss weight	1.0	0.5
Number of epochs	30	7
Optimizer	Adam [72]	Adam [72]

embedding model to learn features invariant to specific forgery techniques.

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

V. RESULTS

The performance is evaluated quantitatively through ablation studies and intra- and cross-dataset evaluations. Qualitative analysis is provided through t-SNE visualization of the learned feature embeddings.

A. Quantitative Results

Table IV presents a comprehensive evaluation of face forgery detection methods under standardized DeepfakeBench [7] settings, with all models trained on the FF++ (HQ) dataset. The intra-dataset results in this table follow scenario 2 (full dataset training) protocol, while the cross-dataset results reflect scenario 3 (cross-dataset evaluation). We report both intra-class and cross-dataset performance using AUC and ACC metrics. The AUC and ACC results in Table IV are largely consistent in overall ranking within datasets. Our performance advantages are consistent across both metrics: datasets where Trident leads in AUC also show corresponding gains in ACC. This indicates that the improvements are not an artifact of threshold choice or class imbalance. Unless otherwise marked, the results are obtained under the standardized DeepfakeBench framework. Methods marked with [†] are included using directly reported numbers from their original papers due to their unavailability within DeepfakeBench. Across most intra-class evaluations, our method achieves the best performance, and in a few cases, it remains highly competitive with the leading baseline. Notably, in cross-dataset scenarios where other methods suffer substantial performance drops, our approach maintains robustness, resulting in a significant improvement over prior work. Additionally, we include runtimes measured on RTX 4090, showing that our method introduces negligible overhead: inference speed is largely dictated by the underlying encoder, ensuring efficiency comparable to baseline models.

1) *Ablation Studies*: Ablation studies were conducted to assess the impact of key components in the *Trident* framework,

TABLE IV
FRAME-LEVEL AUC (↑) AND ACC (↑) SCORES AND MEDIAN INFERENCE TIME (MS) OF FACE FORGERY DETECTORS

Type	Detector	Backbone	Venue	Intra-Dataset Evaluation on FF++ (HQ)										Cross-Dataset Evaluation										Time (ms)				
				FF++		FF-DF		FF-F2F		FF-FS		FF-NT		CDF-v1		CDF-v2		DFD		DFDCP		DFDC			UADFV		WildDF	
				AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC		AUC	ACC	AUC	ACC
Native	MesoNet [73]	Designed CNN	WIFS-18	0.60	80.0	0.66	50.0	0.59	50.0	0.58	50.0	0.57	50.0	0.49	61.6	0.53	65.8	0.54	89.5	0.49	65.7	0.53	52.1	0.58	50.1	0.55	44.5	0.02
	MesoInception [73]	Designed CNN	WIFS-18	0.79	81.4	0.87	66.4	0.83	66.6	0.78	64.8	0.67	60.7	0.69	63.7	0.65	68.0	0.64	82.2	0.71	69.0	0.60	53.5	0.83	65.5	0.55	46.7	0.05
	CNN-Aug [3]	ResNet [74]	CVPR-20	0.84	83.2	0.90	68.9	0.87	68.0	0.86	68.1	0.71	63.9	0.71	68.1	0.67	69.9	0.63	75.8	0.51	63.2	0.60	57.0	0.77	65.3	0.58	44.6	0.20
	Xception [12]	Xception [10]	ICCV-19	0.98	93.2	0.99	87.7	0.99	87.7	0.98	87.1	0.95	85.2	0.75	69.3	0.74	69.5	0.85	80.4	0.73	68.1	0.71	64.2	0.94	79.4	0.72	60.1	0.53
	EfficientNet-B4 [11]	EfficientNet [11]	ICML-19	0.97	92.5	0.97	88.4	0.97	88.6	0.99	89.4	0.95	86.0	0.77	69.4	0.75	66.5	0.81	77.3	0.64	58.4	0.71	63.8	0.92	81.8	0.67	52.9	0.93
Freq.	F3Net [23]	Xception [10]	ECCV-20	0.98	93.8	0.98	92.0	0.99	92.5	0.99	92.5	0.95	88.1	0.75	66.9	0.73	59.1	0.79	68.5	0.70	64.1	0.70	62.6	0.89	80.5	0.65	53.9	0.57
	SPSL [25]	Xception [10]	CVPR-21	0.95	86.8	0.97	68.7	0.97	68.6	0.99	68.8	0.85	67.5	0.80	67.9	0.73	71.0	0.82	88.4	0.72	71.4	0.69	62.4	0.94	64.3	0.68	50.7	0.54
	SRM [75]	Xception [10]	CVPR-21	0.97	91.8	0.98	84.3	0.97	84.6	0.97	84.3	0.94	81.6	0.76	70.7	0.76	72.2	0.84	84.0	0.69	66.4	0.68	62.8	0.88	72.4	0.67	49.6	1.45
Spatial	Capsule [76]	CapsuleNet [77]	ICASSP-19	0.89	85.8	0.90	79.8	0.91	79.7	0.92	80.6	0.81	73.3	0.77	71.9	0.76	73.8	0.71	83.1	0.63	68.0	0.65	59.6	0.95	81.2	0.65	47.7	0.89
	FFD [78]	Xception [10]	CVPR-20	0.96	91.2	0.98	82.6	0.97	82.4	0.99	82.9	0.91	79.8	0.71	66.7	0.69	62.1	0.79	81.9	0.71	71.2	0.71	64.1	0.95	82.9	0.68	48.8	0.53
	CORE [37]	Xception [10]	CVPRW-22	0.97	93.3	0.98	90.7	0.98	90.8	0.98	90.5	0.93	86.3	0.72	66.1	0.74	70.5	0.83	81.7	0.70	67.1	0.70	63.7	0.96	84.5	0.69	55.0	0.53
	RECCE [41]	Custom	CVPR-22	0.97	93.5	0.98	91.5	0.98	92.0	0.98	91.3	0.95	88.6	0.73	67.9	0.74	66.5	0.84	79.6	0.70	65.6	0.68	61.5	0.89	77.6	0.70	54.0	1.30
	UCF [43]	Xception [10]	ICCV-23	0.98	94.6	0.99	92.3	0.99	92.5	0.99	92.6	0.95	88.5	0.81	72.8	0.77	64.9	0.82	75.6	0.69	59.6	0.73	65.8	0.92	81.8	0.71	52.9	1.07
	LSDA [79]	EfficientNet [11]	CVPR-24	0.95	87.6	0.97	72.2	0.98	72.2	0.97	72.0	0.90	70.4	0.87	66.5	0.80	70.8	0.83	88.9	0.78	71.1	0.73	58.6	0.90	62.3	0.69	46.7	0.95
	Effort [80]	CLIP ViT [62]	ICML-25	0.91	85.4	0.99	82.0	0.91	78.2	0.98	81.6	0.75	68.3	0.90	77.8	0.85	78.4	0.90	86.2	0.81	74.1	0.82	69.9	0.97	80.7	0.80	60.5	5.64
	FreqDebias [81] [†]	ResNet [74]	CVPR-25	—	—	—	—	—	—	—	—	—	—	0.88	—	0.84	—	0.87	—	0.82	—	0.74	—	—	—	—	—	—
	UDD [82] [†]	CLIP ViT [62]	AAAI-25	—	—	—	—	—	—	—	—	—	—	0.87	—	—	—	0.91	—	—	—	0.76	—	—	—	—	—	—
	Trident (CNN)	EfficientNet [11]	Ours	0.98	94.6	0.98	90.3	0.98	90.7	0.98	90.2	0.96	86.2	0.76	72.8	0.75	64.9	0.87	87.5	0.84	79.2	0.70	65.8	0.95	81.8	0.75	69.7	0.95
	Trident (ViT)	CLIP ViT [62]	Ours	0.93	86.9	0.99	84.8	0.94	82.2	0.98	84.2	0.79	72.2	0.91	79.0	0.86	78.0	0.92	88.2	0.83	75.2	0.85	75.5	0.97	68.2	0.83	58.2	5.57

Note: The best results are indicated in bold. [†] Directly reported from the original papers due to the unavailability of their code; consequently, some scores are missing. Bold entries indicate the best results.

specifically the CNN backbone, including TL, the GRL, the adversarial loss (Adv), and the detached classification head (DH).

a) Ablation study on temporal triplet sampling: To evaluate the importance of our temporal triplet construction, we compare our proposed aligned sampling strategy with two alternative sampling schemes. The results are reported in Table V. In *aligned sampling (ours)*, the anchor and negative correspond to real and fake frames from the same timestamp, while the positive is sampled from a different timestamp within the same video. This design encourages the model to focus on subtle authenticity cues while controlling for pose, motion, and scene context. In *misaligned sampling*, identity-based triplets are preserved, but the temporal alignment is changed such that the positive and negative frames share the same timestamp instead of the anchor and negative, removing the intended temporal inductive bias. Finally, *random sampling* selects both positive and negative frames randomly from the same video without enforcing any temporal alignment. As shown in Table V, the proposed aligned sampling consistently improves performance across most datasets, demonstrating that enforcing temporal alignment between real and fake frames provides a stronger and more informative training signal for fake detection.

b) Ablation study on all forgery types: Following scenario 2 (full dataset training), all forgery types in FF++ (HQ) are used for both training and testing. Table VI summarizes the results. The baseline model (B) employs the EfficientNet-B4 backbone with a binary classification head. The variants include adding TL and incorporating the GRL (TL+GRL). We observe consistent performance improvements as additional components of the framework are incorporated.

c) Ablation study on deepfakes and NeuralTextures: Under scenario 1 (forgery-specific training), models are trained on only two manipulation types and evaluated on the full FF++ set to probe generalization to unseen forgeries. Table VII presents the results. The methods include the baseline (B), the addition of TL, the incorporation of the GRL (TL+GRL), the use of Adv (TL+Adv), and the inclusion of the DH (TL+GRL+DH). The

TABLE V
ABLATION STUDY RESULTS ON TEMPORAL TRIPLET SAMPLING STRATEGIES IN FRAME-LEVEL AUC (↑): ALIGNED SAMPLING CORRESPONDS TO OUR PROPOSED STRATEGY, WHERE THE ANCHOR AND NEGATIVE ARE TEMPORALLY ALIGNED (REAL VERSUS FAKE FRAMES FROM THE SAME TIMESTAMP), WHILE THE POSITIVE IS SAMPLED FROM A DIFFERENT TIMESTAMP WITHIN THE SAME VIDEO

Dataset	Aligned Sampling (Ours)	Misaligned Sampling	Random Sampling
FaceForensics++	0.93	0.82	0.83
Celeb-DF-v1	0.91	0.83	0.84
Celeb-DF-v2	0.86	0.79	0.79
DFDCP	0.83	0.74	0.77
DFDC	0.85	0.79	0.82
UADFV	0.97	0.95	0.96
DeepFakeDetection	0.92	0.91	0.92

Note: Misaligned Sampling aligns the positive and negative frames instead of the anchor and negative. Random Sampling selects positive and negative frames randomly from the same video. Bold entries indicate the best results.

TABLE VI
ABLATION STUDY ON FF++ (HQ) DATASET (ALL FORGERY TYPES)

Method	AUC (↑)	LogLoss (↓)
Baseline	0.9497	0.2915
TL	0.9613	0.2611
TL+GRL	0.9669	0.2946
TL+GRL+DH	0.9793	0.2456

Note: Bold entries indicate the best results.

results demonstrate consistent performance gains as each component of the proposed framework is progressively introduced.

d) Cross-dataset ablation on DFD (HQ): As defined in scenario 3 (cross-dataset evaluation), models trained on FF++ (HQ) are tested on a completely separate dataset to assess cross-domain robustness. To further assess the impact of our proposed components on generalization, we conducted an ablation study on the DFD (HQ) dataset using models trained on FF++ (HQ). Table VIII presents the results. The *LogLoss* refers to the BCE

TABLE VII
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	0.9652	0.2662
TL+GRL+DH	0.9646	0.2454

Note: Bold entries indicate the best results.

TABLE VIII
CROSS-DATASET ABLATION ON DFD (HQ)
DATASET [TRAINING ON FF++ (HQ)]

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

Note: Bold entries indicate the best results.

TABLE IX
HYPERPARAMETER SENSITIVITY ON THE
DEEPFAKEDETECTION (AUC $\uparrow$) DATASET

Default	β		λ		m	
	0.8	1.2	0.8	1.2	0.8	1.2
0.920	0.929	0.926	0.924	0.926	0.924	0.925

Note: The Default column corresponds to the configuration ($\beta = 1.0$, $\lambda = 1.0$, $m = 1.0$). Bold entries indicate the best results.

loss used for the classification task in both tables. We observe that the generalization performance improves progressively as each component of our framework is added. Notably, the effect of the DH becomes more apparent in this cross-dataset setting compared to the one presented in Table VII, where decoupling the classifier from backbone feature learning leads to more transferable representations and improved robustness under domain shift.

e) Hyperparameter sensitivity analysis: To assess the sensitivity of *Trident* (ViT) to its key training hyperparameters, we vary the forgery loss weight (β), GRL scaling factor (λ), and triplet margin (m) individually while keeping the remaining two at their default value of 1.0. We run these experiments on the DeepFakeDetection dataset. The results are summarized in Table IX. Across all runs, the AUC scores remain stable under both increases and decreases of each hyperparameter, indicating that the framework is robust to moderate perturbations and not overly reliant on precise hyperparameter tuning.

f) Full fine-tuning versus PEFT (BitFit): Table X compares two fine-tuning strategies for the CLIP ViT-L/14 backbone within the *Trident* framework: full fine-tuning of all backbone parameters versus BitFit [6], a parameter-efficient method that updates only bias terms. BitFit yields higher AUC on six of seven datasets, while full fine-tuning outperforms it only on DFDCP. These results indicate that restricting updates to bias parameters is sufficient for effectively adapting large vision-language features to the forgery detection task.

TABLE X
AUC ($\uparrow$) COMPARISON OF FULL FINE-TUNING
VERSUS BITFIT FINE-TUNING OF CLIP ViT-L/14
WITHIN *TRIDENT*

Dataset	Full Fine-Tune	BitFit
FaceForensics++	0.789	0.930
DeepFakeDetection	0.901	0.920
Celeb-DF-v1	0.903	0.910
Celeb-DF-v2	0.851	0.860
DFDCP	0.882	0.830
DFDC	0.849	0.850
UADFV	0.927	0.970

Note: Both variants share identical architecture, training data, and hyperparameters; only the set of updated parameters differs. Bold indicates the higher value per dataset. Bold entries indicate the best results.

TABLE XI
MULTISOURCE TRAINING RESULTS FOR *TRIDENT* (ViT) TRAINED
JOINTLY ON FF++ (HQ) AND CELEB-DF-v2, COMPARED WITH THE
SINGLE-SOURCE FF++ (HQ) BASELINE, EVALUATED ON
EIGHT DATASETS

Test Dataset	Single-source (FF++)		Multi-source	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)
FaceForensics++	86.9	93.0	87.9	93.2
Celeb-DF-v2	78.0	86.1	75.5	96.9
Celeb-DF-v1	79.0	91.2	92.0	99.3
DeepFakeDetection	88.2	92.0	86.7	91.7
DFDC	75.5	84.8	73.7	81.6
DFDCP	75.2	83.0	73.3	85.4
UADFV	68.2	96.7	72.5	96.7
WildDeepfake	58.2	82.7	58.2	83.1

Note: Bold entries indicate the best results.

g) Multisource training generalization: To examine whether the proposed framework remains stable when trained on heterogeneous sources rather than a single dataset, we jointly train *Trident* (ViT) on FF++ (HQ) and Celeb-DF-v2 and evaluate it on the two source datasets and six unseen targets. Training proceeded without instability under the combined source distribution. To assess the robustness gained from heterogeneous training, we compare against the same architecture trained only on FF++ (HQ) (single-source numbers are taken from Table IV). As shown in Table XI, the multisource model retains strong intra-source performance (93.2 AUC on FF++, 96.9 AUC on Celeb-DF-v2) and transfers competitively to unseen domains, including Celeb-DF-v1 (99.3), DeepFakeDetection (91.7), DFDCP (85.4), UADFV (96.7), WildDeepfake (83.1), and DFDC (81.6). Compared with single-source training, AUC improves on Celeb-DF-v1, DFDCP, and WildDeepfake, remains essentially unchanged on FF++ and UADFV, and decreases slightly on DFD and DFDC, suggesting that adding Celeb-DF-v2 reshapes the embedding toward Celeb-style cues at a modest cost on crowd-sourced (DFDC) and high-quality (DFD) benchmarks. The largest gain is on Celeb-DF-v1 (+8.1 AUC), consistent with its distributional proximity to Celeb-DF-v2. These results indicate that the accuracy gains observed under single-source training do not stem from overfitting to one source distribution.

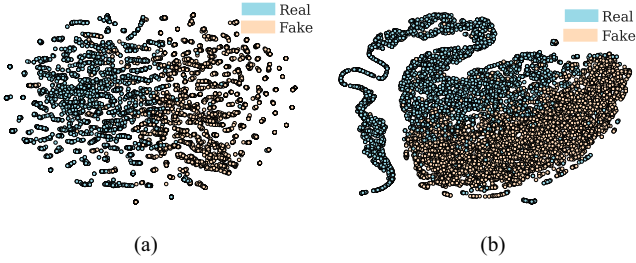


Fig. 2. t-SNE embeddings for models without TL. GRL+DH strengthens real/fake separation but does not yield five distinct forgery clusters. (a) Baseline (B). (b) B+GRL+DH.

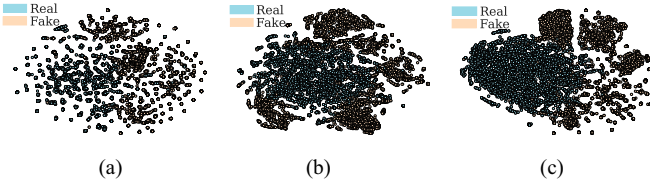


Fig. 3. t-SNE embeddings for TL-based models. TL promotes forgery-type clustering; GRL and DH enhance the separation of real and fake. (a) TL. (b) TL+GRL. (c) TL+GRL+DH.

B. Qualitative Results

The visualizations below are produced under scenario 2 (full dataset training), using the FF++ (HQ) validation split. To qualitatively assess the discriminative capability of the learned embeddings, we employed t-distributed stochastic neighbor embedding (t-SNE) [83] to project high-dimensional features into 2-D space. We present two sets of visualizations: baseline models in Fig. 2 and TL-based models in Fig. 3. Real and fake samples are represented in blue and orange, respectively. All visualizations are produced from the FF++ (HQ) validation split.

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.

We visualize the learned representations using t-SNE to examine how different components shape the embedding space. As shown in Fig. 3(a), using only TL results in some overlap between real and fake samples. Incorporating the GRL (TL+GRL) in Fig. 3(b) leads to a more structured feature space with reduced overlap. Finally, introducing the detached head (TL+GRL+DH) in Fig. 3(c) further enlarges the margin between real and fake samples, leading to a clearer separation in the embedding space.

While the t-SNE visualization in Fig. 3(b) suggests that embeddings may still form clusters corresponding to specific forgery types, we conduct a further analysis to better understand the effect of the domain-adversarial objective. In particular, we extend the visualization by including samples from both the training forgeries and from additional datasets not used during training. The resulting embedding space is shown in Fig. 4. When the GRL is applied, samples from different datasets (e.g.,

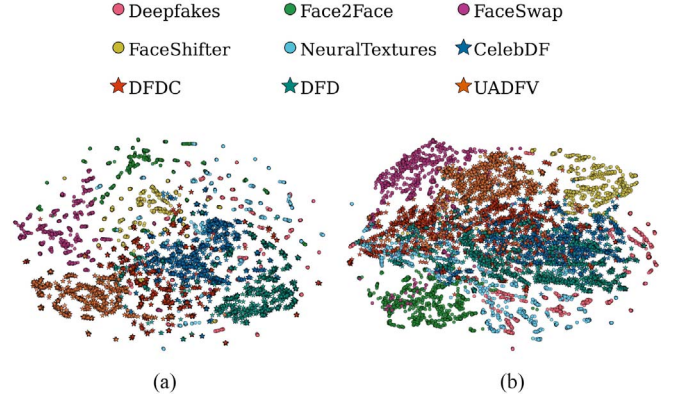


Fig. 4. Extended t-SNE visualization including both training and unseen forgery datasets in TL and TL+GRL models. Although local clusters associated with forgery types are visible, samples from unseen datasets align with these regions instead of forming separate dataset-specific clusters when GRL is applied. (a) TL. (b) TL+GRL.

DFDC, CelebDF, DFD, and UADFV) align with the regions formed during training forgeries rather than forming separate dataset-specific clusters. This indicates that the learned representations capture manipulation-related characteristics that transfer across datasets. This analysis suggests that, although local structures related to forgery types may still emerge in the embedding space, the domain-adversarial training effectively suppresses dataset-specific biases. As a result, the model learns representations that are more robust and transferable to unseen forgery sources.

To further analyze the effect of the domain-adversarial objective, we conduct a probing experiment by freezing the encoder and training a classifier to predict the forgery type from the learned embeddings. When trained on representations obtained with TL, the classifier achieves 76.13% accuracy, indicating that the embedding space retains considerable information about specific manipulation techniques. After introducing GRL (TL+GRL), the accuracy decreases to 70.00%. Although the representation still contains some forgery-type information, the drop in performance suggests that the domain-adversarial objective reduces the model's reliance on technique-specific cues. This behavior aligns with our goal of encouraging the network to focus on more general manipulation artifacts rather than overfitting to particular forgery methods.

We complement the embedding-space view with a spatial one using gradient-weighted class activation mapping (Grad-CAM) [84] on the ViT variant of *Trident* (CLIP ViT-L/14). Gradients of the fake logit at the second-to-last encoder layer are global-average-pooled over the 256 patch tokens; the resulting 16×16 map is upsampled to 224×224 , Gaussian-smoothed ($\sigma = 8$), and overlaid with the JET colormap. As shown in Fig. 5, the activation remains off the face in the genuine sample and localizes to the face regions altered by each manipulation in the five FF++ forgeries as well as in a cross-dataset DFD example, indicating that the detector relies on manipulation-related facial evidence rather than background or compression cues.

VI. DISCUSSION

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

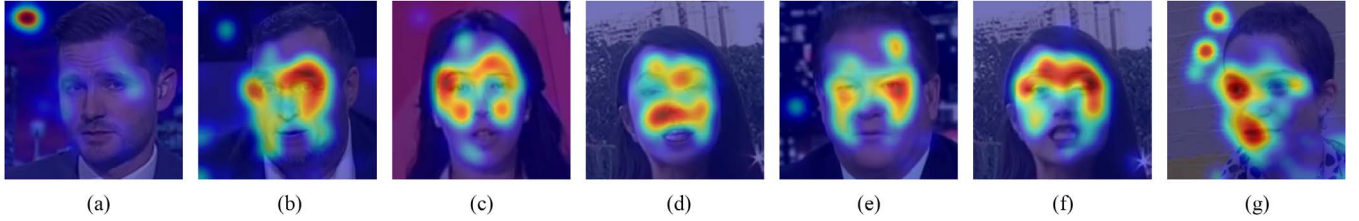


Fig. 5. Grad-CAM heatmaps of *Trident* (ViT) for one real frame, one frame per FF++ forgery type (Deepfakes, Face2Face, FaceSwap, NeuralTextures, FaceShifter), and one cross-dataset frame from DFD. Attention concentrates on the manipulated facial regions. (a) Real. (b) Deepfakes. (c) Face2Face. (d) FaceSwap. (e) NeuralTextures. (f) FaceShifter. (g) DFD.

A. Performance Comparison

The evaluation results in Table IV demonstrate that *Trident* (CNN) achieves an AUC of 0.9793 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 versus Xception’s 0.8163) and DFDCP (0.8403 versus UCF’s 0.7594). The ViT variant outperforms previous methods on multiple benchmarks, achieving the highest AUC scores on Celeb-v1 (0.9115 versus Effort’s 0.9041), Celeb-v2 (0.8606 versus Effort’s 0.8497), DFD (0.9204 versus Effort’s 0.8991), DFDC (0.8476 versus Effort’s 0.8211), UADFV (0.9666 versus Effort’s 0.9659), and WildDeepfake (0.8272 versus Effort’s 0.8004). 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 the impact of each component. From Tables VI and VII, TL 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 features that are forgery-agnostic. The cross-dataset results in Table VIII 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 Figs. 2–4 reveal the models’ discriminative capabilities. Models without TL exhibit 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 separation between real/fake images and distinct clustering for different forgery techniques.

D. CNN Versus 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, consistently improving 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 the task of 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 DFD. 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 than existing approaches, indicating that controlled TL combined with domain-adversarial training offers a promising approach to addressing the fundamental challenges of generalizing forgery detection.

VII. CONCLUSION

We introduce *Trident*, a forgery detection framework that integrates TL and domain-adversarial training to address the

growing sophistication and diversity of face forgery methods. Our approach isolates subtle forgery-specific features by leveraging TL 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 different forgery techniques. Simultaneously, our domain-adversarial training scheme, which incorporates a dedicated forgery discriminator, produces forgery-agnostic embeddings that capture generalizable markers of manipulation.

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 TL, GRLs, 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, multimodal inputs, and advanced metric-learning approaches. Investigating adaptive triplet sampling strategies or dynamic margin selection could refine the embedding space and address the remaining challenges to generalization.

DATA AVAILABILITY STATEMENT

Our code is available at <https://github.com/mhakankara/trident>.

REFERENCES

- [1] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst., NIPS*, 2014, pp. 2672–2680.
- [2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst., NIPS*, vol. 33, 2020, pp. 6840–6851.
- [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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2020, pp. 8695–8704.
- [4] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR*, 2015, pp. 815–823.
- [5] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. 32nd Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37, Lille, France, PMLR, Jul. 7–9, 2015, pp. 1180–1189. Accessed: Jun. 16, 2026. [Online]. Available: <https://proceedings.mlr.press/v37/ganin15.html>
- [6] E., Ben Zaken, Y. Goldberg, and S. Ravfogel, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics* (vol. 2: Short Papers), S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Assoc. for Comput. Linguistics, 2022, pp. 1–9. Accessed: Jun. 16, 2026. [Online]. Available: <https://aclanthology.org/2022.acl-short.1/>
- [7] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "DeepfakeBench: A comprehensive benchmark of deepfake detection," in *Proc. Adv. Neural Inf. Process. Syst., NIPS*, vol. 36, 2023, pp. 4534–4565.
- [8] L. Ma, P. Yang, Y. Xu, Z. Yang, P. Li, and H. Huang, "Deep learning technology for face forgery detection: A survey," *Neurocomputing*, vol. 618, 2025, Art. no. 129055. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231224018265>
- [9] J. W. Seow, M. K. Lim, R. C. Phan, and J. K. Liu, "A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities," *Neurocomputing*, vol. 513, pp. 351–371, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231222012334>
- [10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR*, 2017, pp. 1251–1258.
- [11] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn., ICML*, 2019, pp. 6105–6114.
- [12] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV*, 2019, pp. 1–11.
- [13] L. Li et al., "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2020, pp. 5001–5010.
- [14] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *Proc. 16th Eur. Conf. Comput. Vis., ECCV*, Glasgow, U.K., Part XXVI, Berlin, Heidelberg: Springer-Verlag, Aug. 23–28, 2020, pp. 103–120.
- [15] D. Zhang, R. He, X. Liao, F. Li, J. Chen, and G. Yang, "Face forgery detection based on fine-grained clues and noise inconsistency," *IEEE Trans. Artif. Intell.*, vol. 6, no. 1, pp. 144–158, Jan. 2025.
- [16] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP*, 2019, pp. 8261–8265.
- [17] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2021, pp. 5039–5049.
- [18] Z. Gu et al., "Spatiotemporal inconsistency learning for deepfake video detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2021, pp. 10343–10352.
- [19] I. Amerini, R. Caldelli, and F. Picchioni, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops, ICCVW*, 2019, pp. 1205–1207.
- [20] S. Chhabra, K. Thakral, S. Mittal, M. Vatsa, and R. Singh, "Low-quality deepfake detection via unseen artifacts," *IEEE Trans. Artif. Intell.*, vol. 5, no. 4, pp. 1573–1585, Apr. 2024.
- [21] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi, "Exploiting style latent flows for generalizing deepfake video detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2024, pp. 1133–1143.
- [22] E. Pintelas, I. E. Livieris, and P. E. Pintelas, "Quantization-based 3D-CNNs through circular gradual unfreezing for deepfake detection," *IEEE Trans. Artif. Intell.*, vol. 6, no. 12, pp. 3351–3363, Dec. 2025.
- [23] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. 16th Eur. Conf. Comput. Vis., ECCV*, Glasgow, U.K., Part XII, Berlin, Heidelberg: Springer-Verlag, Aug. 23–28, 2020, pp. 86–103.
- [24] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2021, pp. 6458–6467.

- [25] H. Liu et al., "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2021, pp. 772–781.
- [26] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. 37th Int. Conf. Mach. Learn., ICML, JMLR.org*, 2020, pp. 3247–3258.
- [27] Z. Wang et al., "A timely survey on vision transformer for deepfake detection," 2024, *arXiv:2405.08463*.
- [28] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "ISTVT: Interpretable spatial-temporal video transformer for deepfake detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 1335–1348, 2023.
- [29] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, and R. He, "TALL: Thumbnail layout for deepfake video detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV*, Oct. 2023, pp. 22601–22611.
- [30] J. Wang et al., "M2TR: Multi-scale transformers for deepfake detection," in *Proc. Int. Conf. Multimedia Retrieval, ICMR*, New York, NY, USA: ACM, 2022, pp. 615–623.
- [31] X. Dong et al., "Protecting celebrities from deepfake with identity consistency transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2022, pp. 9468–9478.
- [32] W. Zhuang et al., "UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," in *Proc. 17th Eur. Conf. Comput. Vis., ECCV*, Tel Aviv, Israel, Part V. Berlin, Heidelberg: Springer-Verlag, Oct. 23–27, 2022, pp. 391–407.
- [33] S. A. Khan and H. Dai, "Video transformer for deepfake detection with incremental learning," in *Proc. 29th ACM Int. Conf. Multimedia, MM*, New York, NY, USA: ACM, 2021, pp. 1821–1828.
- [34] D. W. Deressa, H. Mareen, P. Lambert, S. Atnafu, Z. Akhtar, and G. Van Wallendael, "GenConViT: Deepfake video detection using generative convolutional vision transformer," *Appl. Sci.*, vol. 15, no. 12, 2025, Art. no. 6622.
- [35] B. Kaddar, S. A. Fezza, Z. Akhtar, W. Hamidouche, A. Hadid, and J. Serra-Sagristá, "Deepfake detection using spatiotemporal transformer," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 20, no. 11, 2024, Art. no. 345.
- [36] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV*, 2019, pp. 7556–7566.
- [37] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao, "CORE: Consistent representation learning for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, CVPRW*, 2022, pp. 12–21.
- [38] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV*, 2021, pp. 15023–15033.
- [39] Y. Lai, G. Yang, Y. He, Z. Luo, and S. Li, "Selective domain-invariant feature for generalizable deepfake detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP*, 2024, pp. 2335–2339.
- [40] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2022, pp. 18720–18729.
- [41] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2022, pp. 4113–4122.
- [42] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2023, pp. 3994–4004.
- [43] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: Uncovering common features for generalizable deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis., ICCV*, 2023, pp. 22355–22366.
- [44] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2023, pp. 24480–24489.
- [45] Y. Lai et al., "GM-DF: Generalized multi-scenario deepfake detection," in *Proc. 33rd ACM Int. Conf. Multimedia, MM*, New York, NY, USA: ACM, 2025, pp. 4300–4309. [Online]. Available: <https://doi.org/10.1145/3746027.3755386>
- [46] S. Das, M. Kolahdouzi, and A. Etemad, "Kernel space conditional distribution alignment for improving group fairness in deepfake detection," *Trans. Mach. Learn. Res.*, 2025. [Online]. Available: <https://openreview.net/forum?id=68Lv6v4N9J>
- [47] S. Fung, X. Lu, C. Zhang, and C.-T. Li, "DeepfakeUCL: Deepfake detection via unsupervised contrastive learning," in *Proc. Int. Joint Conf. Neural Netw., IJCNN*, 2021, pp. 1–8.
- [48] Y. Xu, K. Raja, and M. Pedersen, "Supervised contrastive learning for generalizable and explainable deepfakes detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops, WACVW*, 2022, pp. 379–389.
- [49] Z. Gu, T. Yao, Y. Chen, S. Ding, and L. Ma, "Hierarchical contrastive inconsistency learning for deepfake video detection," in *Proc. 17th Eur. Conf. Comput. Vis., ECCV*, Tel Aviv, Israel, Part XII. Berlin, Heidelberg: Springer-Verlag, Oct. 23–27, 2022, pp. 596–613.
- [50] B. Liang, Z. Wang, B. Huang, Q. Zou, Q. Wang, and J. Liang, "Depth map guided triplet network for deepfake face detection," *Neural Netw.*, vol. 159, pp. 34–42, 2023.
- [51] A. Kumar, A. Bhavsar, and R. Verma, "Detecting deepfakes with metric learning," in *Proc. 8th Int. Workshop Biometrics Forensics, IWBf*, 2020, pp. 1–6.
- [52] H. Wu, J. Zhou, and S. Zhang, "Generalizable synthetic image detection via language-guided contrastive learning," *IEEE Trans. Artif. Intell.*, vol. 7, no. 6, pp. 3485–3496, Jun. 2026.
- [53] A. L. Pellicier, Y. Li, and P. Angelov, "PUDD: Towards robust multi-modal prototype-based deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, CVPRW*, Jun. 2024, pp. 3809–3817.
- [54] Y. Li and P. Angelov, "Detecting cross-domain deepfake videos with contrastive prototype learning," in *Proc. Int. Joint Conf. Neural Netw.*, Mar. 2025, pp. 1–8.
- [55] Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, KDD*, 2018, pp. 849–857.
- [56] Y. Lai, Z. Yu, J. Wang, L. Shen, Y. Xu, and X. Cao, "Agent4FaceForgery: Multi-agent LLM framework for realistic face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2026, pp. 14073–14083.
- [57] Y. Lai, Z. Yu, J. Wang, L. Shen, Y. Xu, and X. Cao, "ForensicZip: More tokens are better but not necessary in forensic vision-language models," 2024, *arXiv:2603.12208*.
- [58] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Jun. 2009.
- [59] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., CVPR*, vol. 1. Piscataway, NJ, USA: IEEE, 2005, pp. 539–546.
- [60] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR*, 2016, pp. 4004–4012.
- [61] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst., NIPS*, vol. 6, 1993, pp. 737–744.
- [62] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn., ICML*, 2021, pp. 8748–8763.
- [63] Deepfakes Community, "Deepfakes," 2018. [Online]. Available: <https://github.com/deepfakes/faceswap>
- [64] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR*, 2016, pp. 2387–2395.
- [65] M. Kowalski, "FaceSwap," 2018. [Online]. Available: <https://github.com/MarekKowalski/FaceSwap/>
- [66] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, 2019, Art. no. 66.
- [67] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2020, pp. 3207–3216.
- [68] J. Google, "DeepFake detection dataset," 2019. [Online]. Available: <https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html>
- [69] B. Dolhansky et al., "The deepfake detection challenge (DFDC) dataset," 2020, *arXiv:2006.07397*.
- [70] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

- [71] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2382–2390.
- [72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015, *arXiv:1412.6980*.
- [73] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur., WIFS*, Piscataway, NJ, USA: IEEE, 2018, pp. 1–7.
- [74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR*, 2016, pp. 770–778.
- [75] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2021, pp. 16317–16326.
- [76] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., ICASSP*, 2019, pp. 2307–2311.
- [77] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst., NIPS*, vol. 30, 2017, pp. 3859–3869.
- [78] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2020, pp. 5781–5790.
- [79] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu, "Transcending forgery specificity with latent space augmentation for generalizable deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., CVPR*, 2024, pp. 8984–8994.
- [80] Z. Yan et al., "Orthogonal subspace decomposition for generalizable AI-generated image detection," in *Proc. Int. Conf. Mach. Learn., ICML*, 2025, pp. 1–21. <https://openreview.net/pdf?id=GFpjO8S8Po>.
- [81] H. Kashiani, N. A. Talemi, and F. Afghah, "FreqDebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 8775–8785.
- [82] X. Fu, Z. Yan, T. Yao, S. Chen, and X. Li, "Exploring unbiased deepfake detection via token-level shuffling and mixing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 3, 2025, pp. 3040–3048. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/32312>
- [83] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
- [84] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis., ICCV*, 2017, pp. 618–626.



Mustafa Hakan Kara received the B.Sc. and M.S. degrees in computer engineering in 2022 and 2025, respectively, from Bilkent University, Ankara, Türkiye, where he is currently working toward the Ph.D. degree in computer engineering.

He is currently a Machine Learning Engineer with HAVELSAN, Ankara. His research interests include visual disinformation, with a focus on face forgery detection, computational social science, and visual perception for autonomous systems.



Aysegul Dundar received the B.Sc. degree in electrical and electronics engineering from Bogazici University, Istanbul, Türkiye, in 2011, and the Ph.D. degree in machine learning from Purdue University, West Lafayette, IN, USA, in 2016.

She is currently an Assistant Professor in computer science with Bilkent University, Ankara, Türkiye. She was a Senior Research Scientist with NVIDIA, Santa Clara, CA, USA. Her research focuses on domain adaptation, image segmentation, and generative models for image synthesis and

manipulation.

Dr. Dundar won first place in the Domain Adaptation for Semantic Segmentation Competition in the Workshop on Autonomous Vehicle Challenge at CVPR in 2018.



Uğur Güdükbay (Senior Member, IEEE) received the B.S. degree in computer engineering from the Middle East Technical University, Ankara, Türkiye, in 1987, and the M.Sc. and Ph.D. degrees in computer engineering and information science from Bilkent University, Ankara, in 1989 and 1994, respectively.

He is currently a Professor with the Department of Computer Engineering, Bilkent University. He conducted research as a Postdoctoral Fellow at the Human Modeling and Simulation Laboratory, University of Pennsylvania, Philadelphia, PA, USA. His research interests include deep learning for human modeling and animation, conversational virtual agents, personality and emotion synthesis, crowd simulation, rendering, and visualization.

Dr. Güdükbay is a Senior Member of ACM.