Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Towards Generalizable Face Forgery Detection via Multi-Granularity Fusion

Mingjie Zhao
Hong Kong Baptist University
Yiu-ming Cheung
ymc@comp.hkbu.edu.hk
Hong Kong Baptist University https://orcid.org/0000-0001-7629-4648
Guilin Pang
Hong Kong Baptist University

Research Article

Keywords: Face Forgery Detection, Deepfake Detection, CLIP Adaptation, Generalization

Posted Date: July 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10377594/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Towards Generalizable Face Forgery Detection via Multi-Granularity Fusion

Mingjie Zhao[1], Yiu-ming Cheung[1*], Guilin Pang[1]

[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China.

*Corresponding author. E-mail: ymc@comp.hkbu.edu.hk;
Contributing authors: mjzhao@comp.hkbu.edu.hk; csglpang@comp.hkbu.edu.hk;

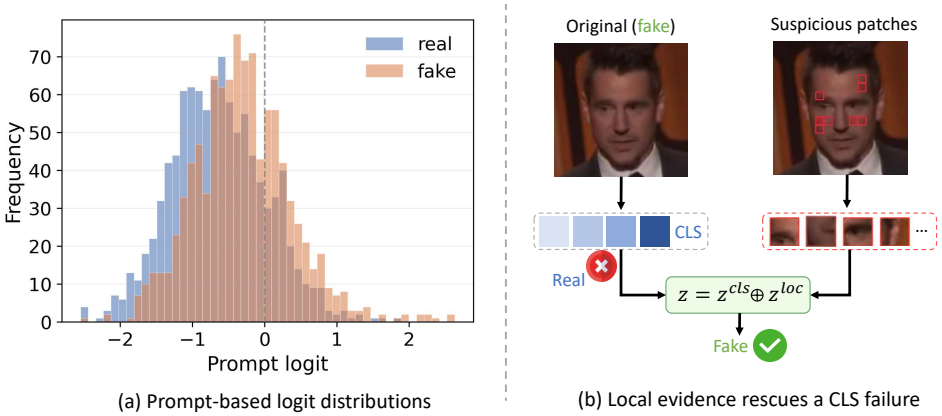
Abstract

Continuously evolving DeepFake techniques threaten visual media authenticity, thus making Face Forgery Detection (FFD) essential for trustworthy digital media. In the literature, despite existing detectors achieving promising intra-dataset performance, they often degrade substantially on unseen DeepFake methods or datasets, which has motivated recent Contrastive Language-Image Pre-training (CLIP)-based detectors to leverage transferable vision-language semantics for more generalizable forgery detection. Nevertheless, such semantic transferability does not ensure reliable forgery perception, since real and fake faces share similar high-level semantics while discriminative cues are often sparse, local, and fine-grained, making it challenging to preserve transferable semantics and capture localized artifacts simultaneously. To address this issue, we therefore propose a Semantic-Consistent Multi-Granularity Fusion framework (SemFusion), which aims to preserve CLIP's transferable semantic structure while enhancing localized forgery perception. Specifically, we introduce Semantic Consistency Regularization (SCR), which uses the frozen CLIP text space as a semantic reference to regularize the adapted visual representation, further stabilizing CLIP's transferable semantics during visual adaptation. Meanwhile, we develop Multi-level Patch Evidence (MPE) learning to mine suspicious local regions from multi-layer visual patch tokens, whose local evidence is fused with global evidence at the logit level for complementary decision making. Extensive experiments demonstrate that SemFusion outperforms the state-of-the-art methods.

Keywords: Face Forgery Detection, Deepfake Detection, CLIP Adaptation, Generalization

1 Introduction

Recent advances in DeepFake techniques, such as face reenactment (Thies, Zollhofer, Stamminger, Theobalt, & Niessner, 2016), face swapping (Q. Li, Wang, Xu, Sun, & Yang, 2024), and GAN- or diffusion-based face synthesis (Karras, Laine, & Aila, 2019; Rombach, Blattmann, Lorenz, Esser, & Ommer, 2022), have made manipulated facial media increasingly realistic and accessible, raising concerns about visual media authenticity. Reliable Face Forgery Detection (FFD) has therefore become an important task for trustworthy digital media and multimedia forensics. Although many FFD methods have achieved promising performance under intra-dataset evaluation, recent benchmarks show that their performance often drops substantially on unseen datasets or manipulation methods (Yan, Zhang, Yuan, Lyu, & Wu, 2023). This degradation is mainly caused



(a) Prompt-based logit distributions

(b) Local evidence rescues a CLS failure

Fig. 1: Visualization of prompt separability and local patch evidence. (a) The prompt logit is computed by frozen CLIP as the similarity margin between fake and real textual prompts, the overlapped real/fake distributions indicate limited separability of direct prompt-based scoring. (b) For a representative fake sample, the global CLS evidence predicts real, while incorporating suspicious local patch evidence corrects the prediction to fake, showing the complementarity between global and local evidence.

by the tendency of detectors to exploit source-specific visual regularities, such as manipulation-specific artifacts, compression statistics, or generator shortcuts, which may not persist across domains (Cheng et al., 2024). Therefore, FFD methods need to generalize beyond the training distribution when facing unseen datasets and manipulation methods.

To this end, existing methods have explored different forensic cues and learning strategies to improve the generalization of FFD. For example, papers like (L. Li et al., 2020; Nirkin, Wolf, Keller, & Hassner, 2022) exploit spatial forensic cues, such as blending boundaries or inconsistencies between manipulated faces and their surrounding context. Besides that, papers like (Cao et al., 2022; Liu et al., 2021; Qian, Yin, Sheng, Chen, & Shao, 2020) mine frequency-domain clues or reconstruction inconsistencies to capture less visible manipulation traces. However, forensic cues captured from specific manipulation pipelines may not remain stable when encountering unseen forgery methods or shifted data distributions. To mitigate this limitation, data- and representation-oriented methods synthesize diverse training samples, learn common forgery features, or augment latent distributions to reduce dependence on specific manipulation pipelines (Shiohara & Yamasaki, 2022; Yan, Luo, Lyu, Liu, & Wu, 2024; Yan, Zhang, Fan, & Wu, 2023). Although these methods improve the generalization of FFD to some extent, they still mainly seek transferable cues within visual forgery patterns learned from deepfake data.

To move beyond such source-dependent forensic patterns, large-scale vision-language pre-training offers a new promising way for generalizable FFD by introducing semantic priors learned from massive image-text data. A representative model is Contrastive Language-Image Pre-training (CLIP), which aligns a visual encoder and a text encoder into a shared visual-semantic space through image-text contrastive learning (Radford et al., 2021). Compared with detectors trained only on limited forgery datasets, CLIP provides broad semantic priors and transferable visual representations beyond source-specific forgery data. These properties have motivated recent studies to adapt CLIP to FFD by converting its cross-modal representations into authenticity evidence, including prompt-based image-text matching with real/fake semantics and visual encoder adaptation for forgery prediction (Khan & Dang-Nguyen, 2024; Yan, Wang, et al., 2025).

However, the semantic priors provided by CLIP do not directly translate into reliable forgery evidence, because challenging face forgeries are often characterized by local, subtle, and weakly semantic artifacts rather than clear high-level semantic differences (Y. Chen et al., 2026; Guo, Song, Zhang, Liu, & Liu, 2025). On the textual

side, this makes direct prompt-based scoring unreliable for FFD. Global prompts such as 'a real face' or 'a fake face' can express coarse authenticity concepts, but they cannot stably describe localized cues such as blending boundaries, texture inconsistencies, or manipulation traces around facial components. Although fine-grained descriptions may provide richer information, their quality depends on accurate annotations and may be affected by noisy or hallucinated descriptions (Sun et al., 2025). As shown in Fig. 1(a), the prompt logit computed by frozen CLIP from the similarity margin between fake and real textual prompts produces highly overlapped real/fake distributions, indicating the limited forensic separability of direct prompt-based scoring. On the visual side, although CLIP visual representations are transferable, existing CLIP-based detectors commonly rely on global image-level representations, such as the CLaSs (CLS) token in Vision Transformers (ViT)-based CLIP, for authenticity prediction. (Cui, Li, Luo, Zhou, & Dong, 2025; Yan, Wang, et al., 2025; Zhang et al., 2025). Such global evidence may dilute sparse local artifacts when aggregating holistic facial semantics. As illustrated in Fig. 1(b), a fake sample with normal global facial semantics can be misclassified as real by the CLS-only branch, suggesting that global image-level evidence alone may overlook localized abnormal regions. These observations reveal that adapting CLIP to FFD still faces challenges from both textual prompt-based scoring and global visual representation.

In this paper, we propose a Semantic-Consistent Multi-Granularity Fusion framework (SemFusion) for generalizable FFD. SemFusion is designed to preserve the transferable semantic structure of CLIP while enhancing the perception of localized forgery evidence. Specifically, we introduce Semantic Consistency Regularization (SCR), rather than an additional prompt classifier, to treat the CLIP text space as a semantic reference to regularize the adapted visual representation and stabilize CLIP's transferable semantics during visual adaptation. To capture local forensic evidence beyond global semantics, we develop Multi-level Patch Evidence (MPE) learning, which mines suspicious cues from multi-layer visual patch tokens and complements the global representation with localized forensic information. During inference, global evidence and local patch evidence are fused at the logit level, producing the global-local complementary behavior illustrated in Fig. 1 (b). Extensive experiments demonstrate the effectiveness of SemFusion, where it achieves state-of-the-art generalization performance on multiple unseen forgery benchmarks. The main contributions of this paper are fourfold:

- We propose SemFusion to improve the generalization of FFD by preserving CLIP's transferable semantic structure and complementing global evidence with localized patch-level forgery cues.
- We introduce SCR to treat the CLIP text space as a semantic reference to constrain visual representations, stabilizing CLIP's transferable semantics without relying on textual prompts as direct decision cues.
- We propose MPE to capture sparse and fine-grained forgery cues from multi-layer visual patch tokens, complementing global evidence that may overlook localized abnormal regions.
- We conduct extensive evaluations on five cross-dataset benchmarks and six cross-method benchmarks. SemFusion outperforms the state-of-the-art methods, with an average frame-level AUC of 0.909 and an average video-level AUC of 0.974, respectively.

The remainder of this paper is organized as follows. Section 2 make an overview of related work and preliminaries. Section 3 presents the proposed SemFusion framework in detail. Section 4 reports experimental results and analyses. Finally, Section 5 draws a conclusion.

2 Related Work and Preliminaries

This section reviews existing studies on generalizable FFD (Section 2.1) and CLIP-based FFD (Section 2.2). We also introduce the basic formulation of CLIP to clarify the visual-semantic representation space used in our method.

2.1 Generalizable FFD

FFD has been extensively studied by modeling visual traces left by manipulation pipelines. Early methods mainly focus on artifact-oriented cues. Face X-Ray explicitly models blending boundaries

between manipulated faces and their surrounding context, while other spatial artifact methods exploit face-context discrepancies or warping artifacts caused by face replacement and reenactment (Dang, Liu, Stehouwer, Liu, & Jain, 2020; L. Li et al., 2020; Nirkin et al., 2022). Frequency-based methods further show that forged faces may contain abnormal spectral patterns that are less visible in the RGB domain. Representative approaches such as F3Net, SPSL, and SRM mine frequency-aware clues, phase inconsistency, or high-frequency residuals for forgery detection (Liu et al., 2021; Y. Luo, Zhang, Yan, & Liu, 2021; Qian et al., 2020). In addition, RECCE introduces reconstruction-classification learning to reveal manipulation traces through reconstruction inconsistency, and CORE promotes consistent representation learning for more robust detection (Cao et al., 2022; Ni et al., 2022). These methods demonstrate the importance of forensic artifacts, but the learned cues may still be sensitive to manipulation types, compression levels, and source-domain preprocessing.

To improve generalization to unseen datasets or manipulation methods, recent work has explored data-centric and representation-centric strategies. Data-centric methods attempt to enlarge or diversify the training distribution. SBI (Shiohara & Yamasaki, 2022) synthesizes self-blended images to simulate common face-swapping artifacts, SLADD (L. Chen, Zhang, Song, Liu, & Wang, 2022) introduces adversarially generated self-supervised samples, and ProDet (Cheng et al., 2024) investigates whether deepfake detectors can be trained without real deepfake data by synthesizing pseudo-fake training examples. Representation-centric methods instead aim to reduce the reliance on source-specific artifacts. UCF (Yan, Zhang, Fan, & Wu, 2023) learns common forgery features shared across manipulations, LSDA (Yan, Luo, et al., 2024) augments the latent space to transcend forgery specificity, and CFM (A. Luo et al., 2024) or FoCus (Tian et al., 2024) mines more critical forgery clues beyond a predefined artifact prior. Other methods introduce localized artifact attention, identity-related cues, or temporal/spatiotemporal modeling to improve robustness under more challenging settings (Huang et al., 2023; Nguyen et al., 2024; Yan, Zhao, et al., 2025).

Although these methods have significantly advanced generalizable FFD, most of them still learn transferable evidence mainly from visual forgery datasets, synthetic artifacts, or task-specific augmentation strategies. As a result, the learned detector may remain coupled with the distribution of available training forgeries.

2.2 CLIP and CLIP-based FFD

2.2.1 CLIP Formulation

CLIP (Radford et al., 2021) learns transferable visual-semantic representations from large-scale image-text pairs. It consists of a visual encoder $\mathcal{E}_v$ and a text encoder $\mathcal{E}_t$, which map an image x and a text description t into a shared embedding space. Given a mini-batch of B paired image-text samples $(x_i, t_i)_{i=1}^{B}$, where x_i denotes the i-th image and t_i denotes its paired text description, CLIP extracts the visual embedding $\mathbf{u}_i$ and textual embedding $\mathbf{q}_j$ as

$$\mathbf{u}_i = \frac{\mathcal{E}_v(x_i)}{\|\mathcal{E}_v(x_i)\|_2}, \quad \mathbf{q}_j = \frac{\mathcal{E}_t(t_j)}{\|\mathcal{E}_t(t_j)\|_2}. \tag{1}$$

Based on these normalized embeddings, CLIP measures the compatibility between image x_i and text description t_j using a scaled cosine similarity:

$$\xi_{ij} = \frac{\mathbf{u}_i^\top \mathbf{q}_j}{\tau_{\text{clip}}}, \tag{2}$$

where ξ_{ij} denotes the image-text matching logit and τ_{clip} is the learnable temperature parameter used in CLIP pre-training. CLIP is optimized with a symmetric contrastive objective over all image-text pairs in the mini-batch:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2B} \sum_{i=1}^{B} \left[-\log \frac{\exp(\xi_{ii})}{\sum_{j=1}^{B} \exp(\xi_{ij})} - \log \frac{\exp(\xi_{ii})}{\sum_{j=1}^{B} \exp(\xi_{ji})} \right]. \tag{3}$$

By applying this objective to large-scale image-text pairs, CLIP pulls matched image-text pairs together and pushes mismatched pairs apart, thereby forming a shared visual-semantic embedding space. Owing to such large-scale vision-language pre-training, CLIP provides transferable visual-semantic representations and supports

recognition by comparing visual features with natural-language descriptions. This property has motivated recent face forgery detectors to exploit CLIP either as a transferable visual backbone or as a source of textual semantic guidance.

2.2.2 CLIP-based FFD

Existing CLIP-based detectors can be roughly divided into text-oriented visual-language methods and visual-adaptation-based methods. Text-oriented methods exploit the language branch of CLIP or other large multimodal models to introduce semantic knowledge into forgery detection. CLIPping studies how to adapt vision-language models for universal deepfake detection and shows the benefit of retaining the text branch during CLIP adaptation (Khan & Dang-Nguyen, 2024). MFCLIP introduces a multi-modal fine-grained CLIP framework for diffusion FFD, aiming to enhance detection through fine-grained visual-language information (Zhang et al., 2025). VLFFD further constructs a Face Forgery Text Generator to produce fine-grained forgery descriptions and trains visual-linguistic detectors with both visual and textual supervision (Sun et al., 2025). M2F2-Det combines CLIP-based forgery prompt learning with large language models to produce both detection results and textual explanations (Guo et al., 2025). These methods show that language information can provide useful semantic guidance and improve interpretability for face forensics.

However, using text as a direct authenticity cue or description source also introduces new limitations. Real and fake faces often share the same high-level facial semantics, while the discriminative evidence may correspond to subtle local artifacts, such as blending boundaries, texture inconsistency, or component-level distortions. Therefore, coarse prompts such as 'a real face' and 'a fake face' may not stably describe the actual forensic evidence. Fine-grained textual descriptions provide richer supervision, but they may depend on the quality of annotation or generation pipelines. For example, VLFFD relies on a forgery text generation pipeline to obtain region-aware textual descriptions, and M2F2-Det uses language models mainly for interpretable reasoning and explanation generation (Guo et al., 2025; Sun et al., 2025).

Another line of work adapts the CLIP visual encoder or vision-language model to forgery detection. RepDFD reprograms a visual-language model for general deepfake detection without directly fine-tuning the internal parameters of the foundation model (K. Lin, Lin, Li, Yao, & Li, 2025). Effort proposes orthogonal subspace decomposition for generalizable AI-generated image detection, preserving the principal semantic subspace of CLIP while adapting the residual subspace to forgery-related patterns (Yan, Wang, et al., 2025). Forensics Adapter introduces a task-specific adapter alongside CLIP and guides the adapter to learn blending-boundary traces through dedicated objectives (Cui et al., 2025). By enhancing the interaction between CLIP tokens and adapter tokens, it achieves strong cross-dataset generalization and has become an important baseline for CLIP-based FFD.

Despite their effectiveness, existing visual-adaptation-based CLIP detectors still leave room for improvement. Forensics Adapter relies on forged-region supervision during training, such as manipulation masks or derived boundary maps, to construct masked boundary learning and patch-wise contrastive learning (Cui et al., 2025). This provides explicit local forensic guidance, but such region-level annotations may not always be available in standard or real-world face forgery scenarios. Effort (Yan, Wang, et al., 2025) and other parameter-efficient CLIP adaptation methods (Lai et al., 2026; K. Lin et al., 2025; Yan, Zhao, et al., 2025) preserve useful pretrained knowledge, but their predictions are still mainly based on global image-level representations and do not explicitly integrate multi-level local patch evidence, which limits their ability to effectively detect forgery traces.

3 Proposed Method

This section presents SemFusion for generalizable FFD. We first give the overall learning framework of SemFusion (Section 3.1), then detail SCR (Section 3.2), MPE (Section 3.3), and the final training and inference objectives (Section 3.4).

3.1 Overall Learning Framework of SemFusion

Fig. 2 illustrates the overall framework of SemFusion. Given an input face crop x and its

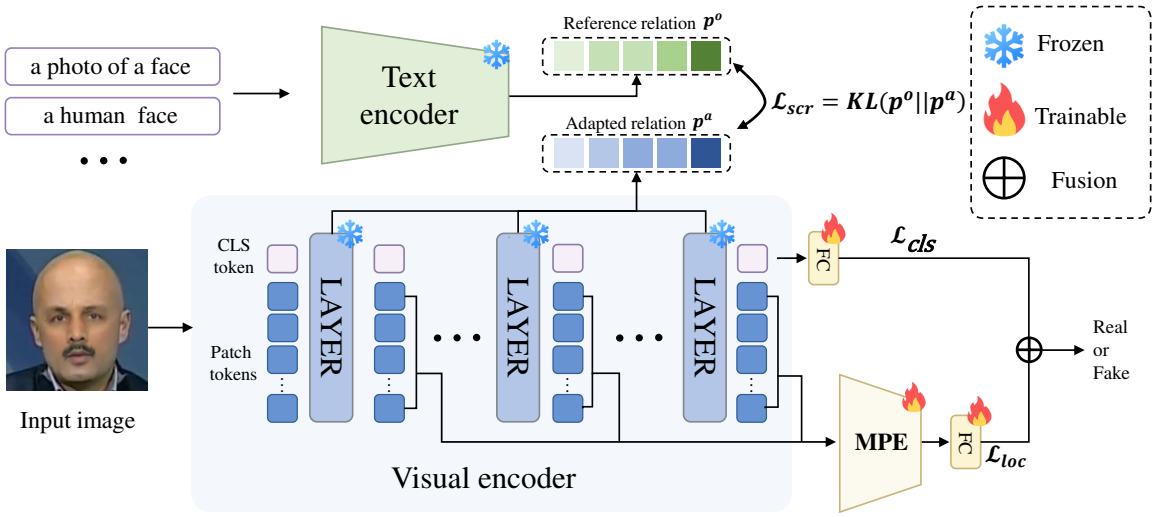


Fig. 2: Overview of SemFusion. The adapted CLIP visual encoder produces the CLS token for global prediction and multi-level patch tokens for local forgery evidence learning. The global branch is supervised by $\mathcal{L}_{\text{cls}}$, while the MPE branch aggregates selected patch tokens to capture sparse local artifacts and is supervised by $\mathcal{L}_{\text{loc}}$. In addition, a frozen text encoder provides semantic anchors from general face prompts, with which the adapted relation distribution p^a is regularized toward the reference relation distribution p^o via $\mathcal{L}_{\text{scr}}$. For simplicity, the visual encoder used to compute p^o is omitted in the diagram. The textual branch is only activated in the training phase.

ground-truth label $y \in \{0, 1\}$, where $y = 1$ and $y = 0$ indicate fake and real images, respectively, FFD aims to predict the authenticity of x. SemFusion uses the adapted CLIP visual encoder for visual evidence extraction and employs the frozen CLIP text space only as a training-time semantic reference. For authenticity prediction, SemFusion builds on a parameter-efficient adaptation of the CLIP visual backbone $\mathcal{G}_v$. Given the input face crop x, the global branch extracts the final CLS-token feature $\mathbf{h}^g$ from $\mathcal{G}_v$ as the image-level representation and feeds it into a lightweight global classification head ϕ_g:

$$\mathbf{z}^g = \phi_g(\mathbf{h}^g). \tag{4}$$

The resulting global logits are denoted as $\mathbf{z}^g = [z_r^g, z_f^g]$, where z_r^g and z_f^g correspond to the real and fake logits, respectively.

To complement the global CLS-token prediction, SemFusion introduces MPE to learn localized patch-level forgery evidence from selected visual layers. Let $\mathcal{S}$ denote the set of selected layers used for local evidence learning. For each selected layer $\ell \in \mathcal{S}$, we denote its patch-token matrix as $\mathbf{V}^\ell$, which contains n patch tokens with visual token dimension d_v. The i-th patch token in this matrix is denoted as $\mathbf{v}_i^\ell$. These selected patch-token matrices $\{\mathbf{V}^\ell\}_{\ell \in \mathcal{S}}$ provide multi-level localized visual cues, from which MPE mines suspicious patch responses and summarizes them into a scalar local fake-evidence logit z^{loc}.

In addition to MPE, SemFusion further introduces SCR as a training-time semantic consistency objective. Therefore, SemFusion is trained with three objectives: the global classification objective for the CLS-token prediction, the local evidence learning objective for MPE, and the semantic consistency objective for SCR. During inference, the training-only SCR branch is discarded, and the final prediction is obtained by fusing the global logits $\mathbf{z}^g$ with the local fake-evidence logit z^{loc}. The following subsections detail SCR, MPE, and the overall training and inference objectives.

3.2 Semantic Consistency Regularization

As discussed in Section 1, directly using textual prompts for real/fake scoring provides limited forensic separability for localized forgery artifacts. Therefore, SemFusion does not use the CLIP text

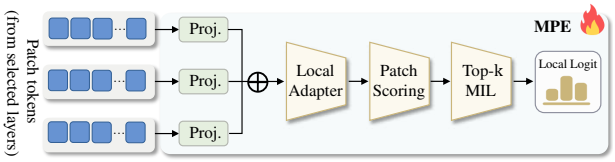


Fig. 3: Details of MPE. Patch tokens are fused and scored, and top-k MIL aggregates suspicious responses into a local fake-evidence logit.

encoder as an additional authenticity classifier. Instead, we use the frozen CLIP text space as a semantic reference to regularize visual adaptation. First, we construct a small bank of general face-related textual descriptions as semantic anchors $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$, where m denotes the number of textual descriptions in the prompt bank.

These descriptions are designed to provide general and domain-agnostic facial semantics, rather than serving as direct real/fake classification prompts. For example, a prompt can be written as *"this is a human face"*, *"a close-up portrait of a face"*, or *"a face with eyes, nose, and mouth"*. Following Eq. (1), each description t_j is encoded by the frozen CLIP text encoder $\mathcal{E}_t$ and normalized as a text anchor $\mathbf{q}_j$. The text anchors are collected as $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_m]$. For an input image, we extract two normalized visual representations, i.e., $\mathbf{u}^a$ from the adapted visual encoder and $\mathbf{u}^o$ from the original frozen visual encoder. Their semantic relation distributions over the text-anchor set are computed as

$$\mathbf{p}^\delta = \mathrm{softmax}\left((\mathbf{u}^\delta)^\top \mathbf{Q}/\tau\right), \quad \delta \in \{a, o\}, \tag{5}$$

where a and o denote the adapted visual encoder and the original frozen visual encoder, respectively, and τ is the temperature parameter. To maintain semantic consistency between the adapted visual representation and the original CLIP semantic space, SCR minimizes the divergence from the frozen reference distribution to the adapted distribution:

$$\mathcal{L}_{\mathrm{scr}} = \mathrm{KL}\left(\mathbf{p}^o \| \mathbf{p}^a\right). \tag{6}$$

This regularization preserves the relative semantic relations between visual representations and frozen text anchors, thereby constraining the adapted visual encoder to remain consistent with the original CLIP semantic space, which helps retain the transferable semantics without overly constraining forgery-related visual adaptation.

However, maintaining semantic consistency alone does not explicitly expose sparse and localized forgery traces, which are often diluted in global image-level representations. This motivates the following MPE learning module to mine localized forensic cues from visual patch tokens.

3.3 Multi-level Patch Evidence Learning

Global CLS-token representations provide image-level semantic evidence for authenticity prediction, but sparse forgery traces may be weakened during global aggregation. To complement the global prediction with localized forensic cues, we introduce MPE, which mines patch-level fake evidence from selected layers of the CLIP-ViT visual encoder. As illustrated in Fig. 3, MPE first projects and fuses multi-level patch tokens, then models local patch interactions, predicts patch-wise fake-evidence scores, and aggregates the most suspicious responses through top-k Multiple Instance Learning (MIL). This design highlights sparse local cues that are most relevant to the fake prediction, avoiding a predefined assumption on where forgery evidence should appear.

Given the patch-token matrix $\mathbf{V}^\ell = [\mathbf{v}_1^\ell, \mathbf{v}_2^\ell, \ldots, \mathbf{v}_n^\ell]$ from a selected visual layer $\ell \in \mathcal{S}$, MPE first maps patch tokens from different layers into a shared local feature space and fuses them at the same patch positions. For the i-th patch, the fused multi-level token is computed as

$$\mathbf{m}_i = \sum_{\ell \in \mathcal{S}} \alpha_\ell \psi_\ell(\mathbf{v}_i^\ell), \tag{7}$$

where $\psi_\ell(\cdot)$ denotes the layer-specific projection for layer ℓ, which maps patch tokens into a shared local adapter dimension, and α_ℓ is the fusion coefficient, which is computed by normalizing the learnable weight β_ℓ of each selected layer:

$$\alpha_\ell = \frac{\exp(\beta_\ell)}{\sum_{\ell' \in \mathcal{S}} \exp(\beta_{\ell'})}, \tag{8}$$

where ℓ' indexes all selected layers in the normalization term. By collecting all fused patch tokens, we obtain the fused multi-level patch-token matrix $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_n]$, which integrates n local

Algorithm 1 The Pseudocode of SemFusion

Require: Training set $\mathcal{D}$, prompt bank $\mathcal{T}$, selected layers $\mathcal{S}$, top-k value k, and loss weights λ_1, λ_2;
Ensure: Trained SemFusion detector;
Encode $\mathcal{T}$ with the frozen text encoder to obtain the text-anchor matrix $\mathbf{Q}$;
TRAINING
for each mini-batch (x, y) sampled from $\mathcal{D}$ **do**
  Extract $\mathbf{h}^g$, $\{\mathbf{V}^\ell\}_{\ell\in\mathcal{S}}$, and $\mathbf{u}^a$ using the adapted visual encoder $\mathcal{G}_v$;
  Compute global logits $\mathbf{z}^g = \phi_g(\mathbf{h}^g)$ and the classification loss $\mathcal{L}_{\text{cls}}$;
  Extract the frozen representation $\mathbf{u}^o$ and compute $\mathcal{L}_{\text{scr}}$ using Eqs. (5)–(6);
  Compute z^{loc} from $\{\mathbf{V}^\ell\}_{\ell\in\mathcal{S}}$ using Eqs. (7)–(11), and obtain $\mathcal{L}_{\text{loc}}$;
  Compute the overall loss $\mathcal{L}$ using Eq. (12) and update all trainable parameters;
end for
INFERENCE
Discard the frozen visual reference encoder and text encoder;
Compute $\mathbf{z}^g$ and z^{loc} using the adapted visual encoder and MPE;
Obtain the final fake probability using Eq. (13).

representations from different visual depths while preserving the patch-wise spatial structure.

To further capture contextual dependencies among local regions, MPE feeds $\mathbf{M}$ into a lightweight transformer encoder $f_{\text{loc}}(\cdot)$ that models interactions among local patch tokens:

$$\mathbf{H}^{\text{loc}} = f_{\text{loc}}(\mathbf{M}). \tag{9}$$

The i-th token of $\mathbf{H}^{\text{loc}}$ is denoted as $\mathbf{h}_i^{\text{loc}}$ and represents the adapted local feature of the i-th patch. For $\mathbf{h}_i^{\text{loc}}$, a local scoring head predicts a patch-level fake-evidence score:

$$s_i = \phi_{\text{loc}}(\mathbf{h}_i^{\text{loc}}), \tag{10}$$

where $\phi_{\text{loc}}(\cdot)$ is a lightweight MLP scoring head that maps each adapted patch feature to a scalar fake-evidence logit. A larger s_i indicates that the corresponding patch contributes stronger evidence to the fake prediction.

To emphasize sparse suspicious evidence and avoid diluting it with irrelevant regions, MPE aggregates the patch scores with a top-k learning strategy, which encourages the local branch to focus on a compact set of suspicious patches rather than all regions. Let $\mathcal{I}_k$ denote the indices of the k largest scores among all n patch scores. The local fake-evidence logit is computed as:

$$z^{\text{loc}} = \frac{1}{k}\sum_{i\in\mathcal{I}_k} s_i. \tag{11}$$

The resulting z^{loc} converts sparse patch-level responses into a local fake-evidence logit, making it compatible with the global CLS-token prediction. In this way, MPE provides a complementary local evidence stream, which can be jointly optimized with the global prediction and later fused with it for final inference.

3.4 Algorithm of SemFusion

With the global semantic prediction, SCR constraint, and local patch evidence defined above, SemFusion is trained by jointly optimizing its corresponding objectives. The global classification loss $\mathcal{L}_{\text{cls}}$ is applied to the two-class global logits $\mathbf{z}^g$, while the local evidence loss $\mathcal{L}_{\text{loc}}$ supervises the top-k aggregated fake-evidence logit z^{loc} using the same label, implemented as a standard binary classification loss (e.g., cross-entropy for binary logistic loss). Together with the semantic consistency regularization $\mathcal{L}_{\text{scr}}$, the overall training objective is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1\mathcal{L}_{\text{scr}} + \lambda_2\mathcal{L}_{\text{loc}}, \tag{12}$$

where λ_1 and λ_2 control the contributions of semantic consistency regularization and local evidence learning, respectively.

Since SCR is formulated as a training-time regularization term and does not produce prediction logits, the frozen CLIP visual reference encoder and the CLIP text encoder are not used at inference. The detector therefore consists only of the adapted visual encoder, the global classification head, and the local evidence branch. Given the global logits $\mathbf{z}^g = [z_r^g, z_f^g]$ and the local fake-evidence logit z^{loc}, SemFusion performs logit-level fusion by adding the local evidence to the fake logit while keeping the real logit unchanged. The final fake probability is computed from the fused two-class logits as

$$p_f = \frac{\exp(z_f^g + z^{\text{loc}})}{\exp(z_r^g) + \exp(z_f^g + z^{\text{loc}})}. \tag{13}$$

This fusion treats the local branch as complementary fake evidence. When suspicious local cues are detected, z^{loc} increases the fake-side evidence; otherwise, the decision mainly relies on the global image-level prediction. The whole algorithm of SemFusion is summarized in Algorithm 1.

4 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness and generalizability of the proposed SemFusion framework. We begin by detailing the experimental setup, including the evaluation datasets, metrics, and implementation details (Section 4.1). Subsequently, we benchmark its generalization capabilities against state-of-the-art detectors under both cross-dataset and cross-method protocols (Section 4.2). To validate the contribution of each proposed component, we perform detailed ablation studies (Section 4.3) and parameter sensitivity analysis (Section 4.4). Furthermore, we assess the model's robustness against common test-time visual corruptions (Section 4.5). Finally, we provide qualitative visual evidence and score distribution analysis to understand the model's decision-making (Section 4.6), delve into the design choices of textual semantic anchors in SCR (Section 4.7), and evaluate the efficiency of SemFusion (Section 4.8).

4.1 Experimental Setup

Datasets. Following previous generalizable FFD protocols (Yan, Luo, et al., 2024; Yan, Wang, et al., 2025; Yan, Zhang, Fan, & Wu, 2023; Yan, Zhang, Yuan, et al., 2023), we trained all variants on the c23 compression version of FF++ (Rossler et al., 2019). For cross-dataset evaluation (Protocol-1), we tested on Celeb-DF-v1, Celeb-DF-v2 (Y. Li, Yang, Sun, Qi, & Lyu, 2020), DeepFakeDetection, DFDC (Dolhansky et al., 2020), and DFDCP (Dolhansky, Howes, Pflaum, Baram, & Ferrer, 2019). For cross-method evaluation (Protocol-2), we evaluated on unseen forgery methods from DF40 (Yan, Yao, et al., 2024), including UniFace, BlendFace, MobileSwap, E4S, FaceDancer, and FSGAN.

Metrics. Our evaluation metrics comprise the Area Under the Receiver Operating Characteristic Curve (AUC) and the Equal Error Rate (EER). We measure the AUC at both the frame and video levels to capture the detection efficacy across different granularities. Alongside AUC, EER is reported to reflect the theoretical lower bound of classification errors.

Implementation Details. We implemented SemFusion based on the DeepfakeBench framework (Yan, Zhang, Yuan, et al., 2023). All experiments were conducted on a single NVIDIA RTX 4090 GPU. CLIP ViT-L/14 (Radford et al., 2021) was adopted as the visual backbone, with input face crops resized to 224×224. Following the standard DeepfakeBench preprocessing pipeline, faces were detected and cropped before being fed into the detector. All experiments were conducted on FF++ c23 as the source training set. During training, we sampled 8 frames per video, while 32 frames were used for testing. For the global branch, SemFusion builds upon a parameter-efficient CLIP-based visual detector, where most pre-trained visual parameters were kept fixed, and only lightweight task-specific parameters were updated to retain CLIP's transferable visual knowledge while learning forgery-related cues (Yan, Wang, et al., 2025). For the local evidence branch, we extracted patch tokens from layers $[8, 16, 24]$ of the CLIP visual encoder. The selected multi-level patch features were fused with learnable layer weights and processed by a lightweight Transformer adapter. The adapter dimension was set to 256, with one Transformer adapter layer and four attention heads. The local evidence was aggregated with top-k pooling, where $k = 16$. During inference, the global fake logit and the local fake-evidence logit are directly added for the final prediction. We trained SemFusion for 10 epochs with a batch size of 32 using the Adam optimizer (Kingma, 2014). The learning rate was set to 2×10^{-4} and the weight decay was 5×10^{-4}. The loss weights are set to $\lambda_1 = 0.2$ and $\lambda_2 = 0.05$. Unless otherwise specified, all reported results were obtained using this configuration.

4.2 Evaluation on Cross-Dataset and Cross-Method Generalization

We evaluated SemFusion under two generalization protocols. In Protocol-1, all methods were trained on FF++ and evaluated on five unseen datasets, including CDF-v1, CDF-v2, DFD, DFDC, and DFDCP. We report **frame-level AUC** to compare cross-dataset generalization. In Protocol-2,

Table 1: Cross-dataset evaluation results (**Protocol-1, Frame-level AUC**). All methods are trained on FF++ and evaluated on other datasets. The best results are indicated in **bold**, and the second-best results are underlined. The reported results are directly collected from their original papers or subsequent comparative studies.

Methods	Venue	CDF-v1	CDF-v2	DFD	DFDC	DFDCP	Avg.
FFD (Dang et al., 2020)	CVPR'20	0.784	0.744	0.802	0.703	0.743	0.755
SRM (Y. Luo et al., 2021)	CVPR'21	0.793	0.755	0.812	0.700	0.741	0.760
SPSL (Liu et al., 2021)	CVPR'21	0.815	0.765	0.812	0.704	0.741	0.767
RECCE (Cao et al., 2022)	CVPR'22	0.760	0.732	0.812	0.713	0.734	0.752
CORE (Ni et al., 2022)	CVPR'22	0.780	0.743	0.802	0.705	0.734	0.753
UCF (Yan, Zhang, Fan, & Wu, 2023)	ICCV'23	0.779	0.753	0.807	0.719	0.759	0.763
ED (Ba et al., 2024)	AAAI'24	0.818	0.864	-	0.721	0.851	-
LSDA (Yan, Luo, et al., 2024)	CVPR'24	0.867	0.830	0.880	0.736	0.815	0.826
ProDet (Cheng et al., 2024)	NeurIPS'24	0.909	0.844	-	0.811	0.724	0.822
FIA-USA (Ma et al., 2025)	NeurIPS'25	0.901	0.867	0.821	-	0.818	-
UDD (Fu, Yan, Yao, Chen, & Li, 2025)	AAAI'25	-	0.860	0.910	0.758	0.856	-
DFA (Shao, Wu, Nie, & Liu, 2025)	IJCV'25	-	0.847	-	0.776	-	-
Effort (Yan, Wang, et al., 2025)	ICML'25	0.926	0.878	0.922	0.822	0.835	0.877
RepDFD (K. Lin et al., 2025)	AAAI'25	-	0.800	-	0.773	**0.906**	-
SRI-Net (Fei, Jia, Huang, Zhang, & Zhou, 2026)	ICLR'26	0.913	0.875	0.893	-	-	-
SemFusion	-	**0.938** (↑1.2%)	**0.917** (↑3.9%)	**0.939** (↑1.7%)	**0.853** (↑3.1%)	0.901 (↓0.5%)	**0.909** (↑3.2%)

Table 2: Benchmarking results for cross-method evaluation (**Protocol-2, Video-level AUC**). All detectors are trained on FF++ and evaluated on unseen forgery methods. The best results are indicated in **bold**, and the second-best results are underlined. The reported results are directly collected from their original papers or subsequent comparative studies.

Methods	Venue	UniFace	BleFace	MobSwap	E4S	FaceDan	FSGAN	Avg.
F3Net (Qian et al., 2020)	ECCV'20	0.809	0.808	0.867	0.494	0.717	0.845	0.757
SPSL (Liu et al., 2021)	CVPR'21	0.747	0.748	0.885	0.514	0.666	0.812	0.729
SRM (Y. Luo et al., 2021)	CVPR'21	0.749	0.704	0.779	0.704	0.659	0.772	0.728
CORE (Ni et al., 2022)	CVPR'22	0.871	0.843	0.959	0.679	0.774	0.958	0.847
RECCE (Cao et al., 2022)	CVPR'22	0.898	0.832	0.925	0.683	0.848	0.949	0.856
SLADD (L. Chen et al., 2022)	CVPR'22	0.878	0.882	0.954	0.765	0.825	0.943	0.875
SBI (Shiohara & Yamasaki, 2022)	CVPR'22	0.724	0.891	0.952	0.750	0.594	0.803	0.786
UCF (Yan, Zhang, Fan, & Wu, 2023)	ICCV'23	0.831	0.827	0.950	0.731	0.862	0.937	0.856
IID (Huang et al., 2023)	CVPR'23	0.839	0.789	0.888	0.766	0.844	0.927	0.842
LSDA (Yan, Luo, et al., 2024)	CVPR'24	0.872	0.875	0.930	0.694	0.721	0.939	0.839
ProDet (Cheng et al., 2024)	NeurIPS'24	0.908	0.929	0.975	0.771	0.747	0.928	0.876
CDFA (Y. Lin et al., 2024)	ECCV'24	0.762	0.756	0.823	0.631	0.803	0.942	0.786
VbSaT (Yan, Zhao, et al., 2025)	CVPR'25	0.960	0.906	0.946	0.980	0.912	0.964	0.945
Effort (Yan, Wang, et al., 2025)	ICML'25	**0.962**	0.873	0.953	0.983	0.926	0.957	0.942
MAP-Mamba (Wang et al., 2026)	TIFs'26	0.815	0.896	0.959	0.779	0.826	-	-
SemFusion	-	**0.962** (-)	**0.955** (↑2.6%)	**0.977** (↑0.2%)	**0.993** (↑1.0%)	**0.971** (↑4.5%)	**0.983** (↑1.9%)	**0.974** (↑2.9%)

detectors were also trained on FF++ but evaluated on unseen forgery methods, including UniFace, BlendFace, MobileSwap, E4S, FaceDancer, and FSGAN. We report video-level AUC to assess cross-method generalization. The compared methods cover classical CNN-based detectors, frequency-aware detectors, recent generalization-oriented methods, and recent foundation-model-based detectors. The reported results were directly collected from their original papers or subsequent comparative studies. Since several frame-based methods only report either **frame-level AUC** or video-level AUC, the included counterparts in Table 1 and Table 2 are not fully consistent.

Table 1 reports the **frame-level AUC** results under Protocol-1. SemFusion achieved the best

Table 3: Ablation study of the proposed components. SCR denotes text semantic regularization, MPE denotes local patch evidence learning, and FU denotes inference-time local-global fusion.

SCR	MPE	FU	CDF-v1		CDF-v2		DFD		DFDC		DFDCP		Avg.	
			AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
-	-	-	0.898	0.178	0.878	0.204	0.913	0.167	0.813	0.262	0.857	0.235	0.872	0.209
✓	-	-	0.895	0.191	0.891	0.188	**0.940**	**0.124**	0.808	0.265	0.864	0.223	0.879	0.198
✓	✓	-	0.924	0.136	0.894	0.177	0.933	0.125	0.849	0.233	0.866	0.225	0.893	0.179
✓	✓	✓	**0.938**	**0.129**	**0.917**	**0.168**	0.939	0.128	**0.853**	**0.230**	**0.901**	**0.185**	**0.909**	**0.168**

average AUC of 0.909 across the five cross-dataset benchmarks, outperforming the strong baseline Effort by 3.2%. Specifically, SemFusion obtained the best results on CDF-v1, CDF-v2, DFD, and DFDC, and achieved the second-best result on DFDCP. The improvement is particularly meaningful on challenging datasets such as CDF-v2 and DFDC, where SemFusion improved the AUC from 0.878 to 0.917 and from 0.822 to 0.853 over the second-best method. These results indicate that SemFusion does not merely fit the source-domain forgery patterns in FF++ but learns more transferable evidence for detecting face forgeries across different data distributions.

Table 2 further reports the cross-method generalization results. SemFusion achieved the best average video-level AUC of 0.974, surpassing both VbSaT and Effort, whose average AUCs are 0.945 and 0.942, respectively. Among the six unseen forgery methods, SemFusion achieved the best performance on BlendFace, MobileSwap, E4S, FaceDancer, and FSGAN, and tied with the best result on UniFace. Compared with Effort, SemFusion brought clear gains on most unseen methods, especially on FaceDancer and BlendFace. This suggests that the proposed global–local fusion is effective not only for cross-dataset transfer, but also for detecting forgeries produced by unseen manipulation techniques. Overall, the two protocols have consistently demonstrated the generalization capability of SemFusion. The gains over Effort under both cross-dataset and cross-method settings suggest that maintaining transferable semantic representations while incorporating local patch evidence provides complementary benefits for generalizable FFD.

4.3 Ablation Study

Table 3 validates the contribution of each component. The baseline with only global CLS classification achieved an average AUC of 0.872. Adding SCR improved the average AUC to 0.879 and reduced EER from 0.209 to 0.198, showing that semantic consistency helps reduce overfitting to source-domain artifacts. Adding MPE further improved the average AUC to 0.893, confirming the value of weakly supervised local evidence. Finally, enabling FU produced the best average AUC and EER, demonstrating that local evidence is useful not only as a training signal but also as complementary inference evidence. Although SCR alone achieved the highest DFD AUC, the full model achieved the best average performance and better results on CDF-v2 and DFDC.

The ablation study has also clarified the different roles of SCR and MPE. SCR improves representation stability but does not explicitly introduce new local evidence at inference. This explains why it improves the average AUC while still showing a mixed effect on DFDC, where low-level artifacts and compression changes may dominate the decision. MPE, on the other hand, directly introduces a local fake-evidence objective, leading to better performance on datasets where sparse artifacts are more informative. The final fusion benefits from both effects: the global branch remains semantically stable, and the local branch can override the global prediction when a small but suspicious region is detected.

4.4 Parameter Sensitivity Analysis

We conducted sensitivity analysis on two key hyperparameters of the local evidence branch: the selected visual layers and the top-k value used for local patch aggregation. All other settings were kept the same as the final SemFusion configuration, and we report frame-level AUC and EER on five cross-dataset benchmarks.

Table 4 compares different layer configurations uniformly sampled from the 24-layer visual encoder. Using two layers, i.e., $[12, 24]$, achieved an average AUC of 0.890 and an average EER

Table 4: Sensitivity to the number of layers used in the Local Adapter. We uniformly sample different numbers of layers from the 24-layer visual encoder.

# Layers	Selected Layers	CDF-v1		CDF-v2		DFD		DFDC		DFDCP		Avg.	
		AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
2	[12, 24]	0.935	0.138	0.897	0.188	0.928	0.140	0.820	0.259	0.872	0.217	0.890	0.188
3	[8, 16, 24]	**0.938**	**0.129**	**0.917**	**0.168**	**0.939**	**0.128**	**0.853**	**0.230**	**0.901**	**0.185**	**0.909**	**0.168**
4	[6, 12, 18, 24]	0.932	0.139	0.894	0.192	0.934	0.131	0.830	0.247	0.883	0.206	0.894	0.183

Table 5: Sensitivity to the top-k value in the proposed Local Adapter MIL

top-k	CDF-v1		CDF-v2		DFD		DFDC		DFDCP		Avg.	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
8	**0.940**	**0.124**	0.916	0.169	0.939	0.127	0.851	0.230	0.898	0.189	0.909	0.168
16	0.938	0.129	**0.917**	**0.168**	**0.939**	**0.128**	0.853	0.230	**0.901**	**0.185**	**0.909**	**0.168**
32	0.936	0.133	0.916	0.168	0.939	0.127	**0.854**	**0.228**	0.899	0.188	0.909	0.169
64	0.934	0.134	0.915	0.168	0.937	0.129	0.854	0.228	0.894	0.193	0.907	0.170

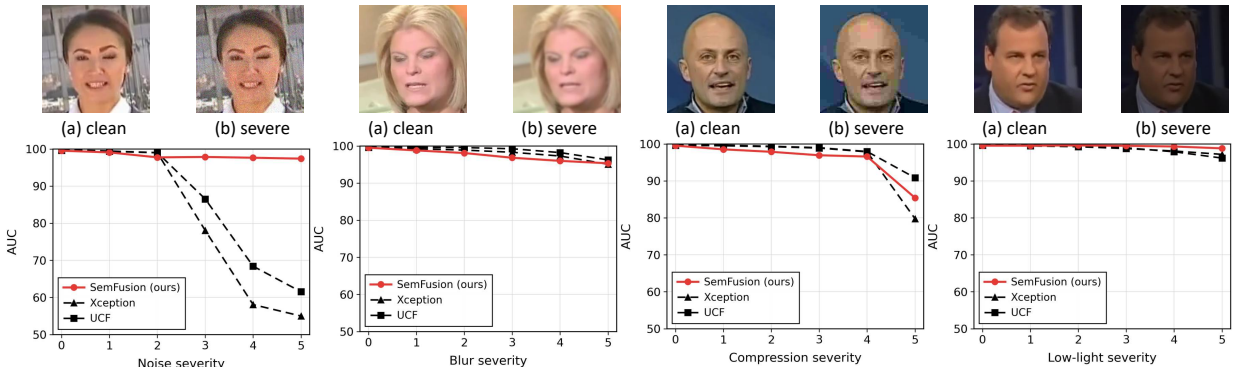


Fig. 4: Robustness Analysis. Our method is compared with Xception and UCF across five levels of four particular types of perturbations in video-level AUC.

of 0.188. Introducing one additional intermediate layer, i.e., $[8, 16, 24]$, improved the average AUC to 0.909 and reduced the average EER to 0.168, giving the best overall performance. This suggests that combining patch representations from multiple visual depths is beneficial for local forgery evidence learning. However, increasing the number of layers to four, i.e., $[6, 12, 18, 24]$, did not further improve the results, indicating that simply adding more layers may introduce redundant or less discriminative local responses.

Table 5 studies the influence of the top-k value. The performance remained stable when k varied from 8 to 64, showing that the local evidence aggregation is not overly sensitive to this hyperparameter. Among these settings, $k = 16$ achieved the best or near-best results on most datasets and provided the most balanced AUC/EER trade-off. Therefore, we adopted $[8, 16, 24]$ and top-$k = 16$ as the default configuration in SemFusion.

4.5 Robustness to Corruptions

To examine whether SemFusion remains reliable under common visual degradations, we conduct a test-time corruption study on FF++. We compare SemFusion with two representative detectors from DeepfakeBench: Xception (Chollet, 2017), a widely used spatial CNN baseline, and UCF (Yan, Zhang, Fan, & Wu, 2023), a generalization-oriented detector based on uncertainty-guided

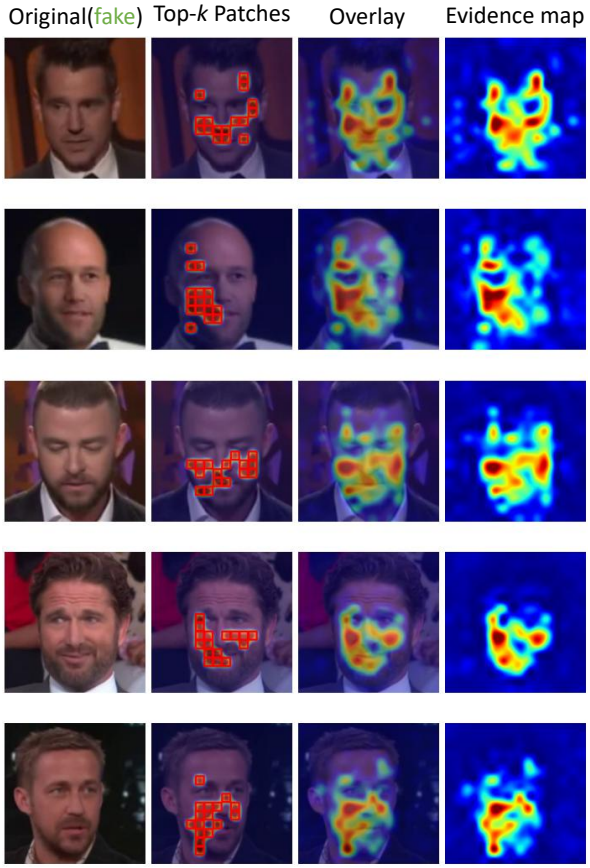


Fig. 5: Visualization of local patch evidence on CDF-v2. All samples are ground-truth fake frames where the global CLS branch predicts real, but the final global-local fusion predicts fake.

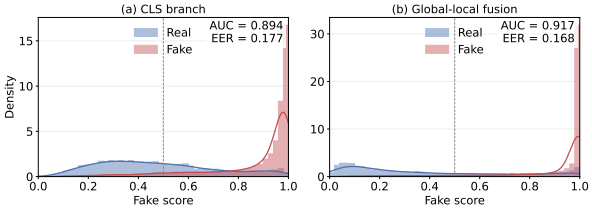


Fig. 6: Score distribution comparison between the CLS branch and global–local fusion on CDF-v2. The dashed vertical line denotes the decision threshold of 0.5.

common feature learning. All methods are evaluated using their corresponding pretrained checkpoints under the same testing protocol, without corruption-specific retraining, fine-tuning, or threshold adaptation. we evaluated four common corruption families, including Gaussian noise, Gaussian blur, JPEG compression, and low-light degradation. For each corruption type, we use a unified severity index from 0 to 5 on the x-axis of Fig. 4, where severity 0 denotes clean samples and severity levels 1–5 indicate progressively stronger corruptions. Specifically, the five non-clean severity levels correspond to noise standard deviations $\sigma = \{2, 5, 10, 15, 20\}$, blur kernel sizes $k = \{3, 5, 7, 9, 11\}$, JPEG quality factors $Q = \{90, 70, 50, 30, 10\}$, and brightness factors $\{0.9, 0.8, 0.7, 0.6, 0.5\}$ for low-light degradation. Note that lower JPEG quality and lower brightness factor indicate stronger degradation. All corruptions are applied only at test time to the loaded RGB face crops before model normalization, while the training procedure and model parameters remain unchanged.

As shown in Fig. 4, all methods achieve comparable clean performance on FF++, and the key difference lies in their degradation trends under increasing corruption severity. SemFusion shows substantially better stability under Gaussian noise, where Xception and UCF degrade sharply at high severity levels. It also remains robust under low-light degradation, with only marginal AUC variation. For Gaussian blur, the three methods show relatively small performance changes, suggesting that moderate blur does not strongly affect in-distribution detection. Under JPEG compression, all methods experience degradation at severe levels, and SemFusion remains competitive, although it is not uniformly superior at the strongest compression level. Overall, the results indicate that SemFusion is less sensitive to common visual corruptions, especially noise and illumination changes, while severe compression remains a challenging case due to the suppression of local forgery evidence.

4.6 Local Evidence Visualization

To better understand the role of the local evidence branch, we provide both qualitative evidence visualization and score distribution analysis on CDF-v2. We first visualized representative fake frames selected from local rescue cases, where the global CLS branch misclassified the fake frame as real, while the final global–local fusion correctly classified it as fake. This setting directly

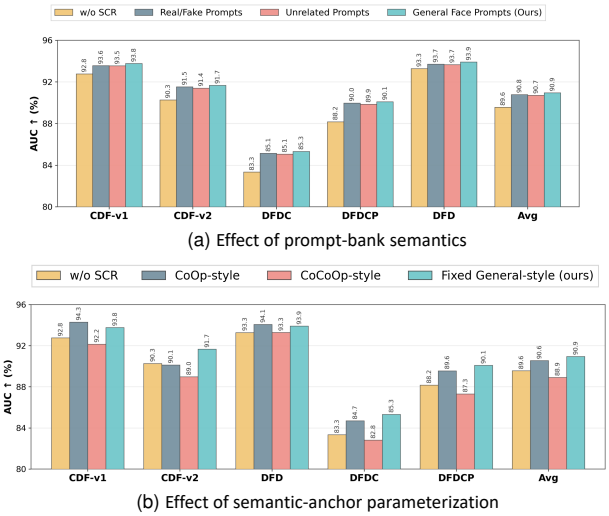


Fig. 7: Analysis of semantic anchors in SCR. (a) Effect of prompt-bank semantics. Different prompt banks are used as SCR anchors while keeping the SemFusion architecture unchanged. (b) Effect of semantic-anchor parameterization.

examines whether local evidence can provide complementary cues beyond the global representation. For each sample, Fig. 5 shows the original face crop, the suspicious patches selected by the local branch, the evidence overlay, and the standalone local evidence map. The evidence map visualizes the relative local responses used by SemFusion for forgery prediction.

As shown in Fig. 5, the highlighted patches and high-response regions are mainly located on facial areas rather than the background. They often appear around local regions such as the mouth, cheeks, eyes, and facial boundaries, where subtle artifacts or blending inconsistencies are likely to occur. Since these examples correspond to cases where the global branch fails but the fused prediction succeeds, the visualization suggests that the local branch captures complementary forgery evidence that may be diluted in global representations. It should be noted that these maps are not pixel-level manipulation masks, but qualitative evidence visualizations of the detector's local responses.

To further verify whether the local evidence improves the overall decision distribution, we compared the fake-score distributions of the CLS branch and the final global–local fusion on the CDF-v2 test set. As shown in Fig. 6, the CLS branch already assigns high scores to many fake samples, but there remains a noticeable overlap between the real and fake distributions. After incorporating local evidence, the fake distribution becomes more concentrated near high fake scores, while the overlap between real and fake samples is reduced. Correspondingly, the frame-level AUC improved from 0.894 to 0.917, and the EER decreased from 0.177 to 0.168. These results have demonstrated that local evidence not only explains individual rescue cases but also improves the separability of real and fake samples.

4.7 Analysis of Semantic Anchors in SCR

To better understand the role of text anchors in SCR, we analyzed semantic anchors from two perspectives: the semantic content of the prompt bank and the parameterization of the anchors. Specifically, we first studied whether SCR depends on face-related prompt semantics, and then compared fixed anchors with learnable and sample-specific anchor variants.

Effect of Prompt Bank Design. We first studied how the design of the prompt bank affects SCR. All variants used the same SemFusion architecture, and only the text anchors used in SCR were changed. Specifically, we compared three types of prompt banks. The first used authenticity-oriented prompts, such as *"a real face"* and *"a fake face"*, which explicitly describe real/fake semantics. The second used unrelated natural-language prompts, such as object- or scene-related descriptions like *"a photo of a mountain"*, to examine whether SCR benefits only from face-related textual semantics. The third used general face prompts, such as *"this is a human face"* and *"a close-up portrait of a face"*, which describe domain-agnostic facial semantics without directly encoding authenticity labels. In addition, we included a variant without SCR as the baseline.

As shown in Fig. 7(a), all prompt-anchor variants improved over the model without SCR. This observation indicates that the benefit of SCR does not simply come from a particular prompt wording. Instead, the KL constraint over frozen text-anchor relations acts as a semantic geometry preservation mechanism, which discourages the adapted visual representation from drifting excessively from the original CLIP space.

Table 6: Efficiency comparison of representative detectors. FPS denotes inference throughput, measured with pure model forward and excluding data loading and preprocessing.

Method	Backbone	Inf. Params (M)	Trainable (M)	FLOPs (G)	FPS
Xception	Xception	21.86	21.86	5.98	492.37
UCF	Xception	46.84	46.84	6.07	224.87
Effort	CLIP ViT-L/14	304.17	0.20	81.18	93.95
SemFusion	CLIP ViT-L/14	305.52	1.55	81.46	89.29

Interestingly, even unrelated prompts brought clear improvements over disabling SCR. This suggests that natural-language text embeddings, although not face-related, can still define a stable reference basis in the CLIP text space. Matching the adapted visual representation to the frozen visual representation under this reference basis provides a useful regularization effect. Nevertheless, the proposed general face prompts achieved the best average AUC. Compared with unrelated prompts, face-related prompts provide a more task-relevant semantic reference while remaining general and domain-agnostic. Therefore, we adopt general face prompts as the default SCR anchors, as they offer a better balance between semantic relevance and simplicity.

Fixed vs. Learnable Semantic Anchors. We further investigated whether learnable semantic anchors can improve SCR. Specifically, we compared the proposed fixed general face anchors with CoOp-style (Zhou, Yang, Loy, & Liu, 2022b) learnable anchors and CoCoOp-style (Zhou, Yang, Loy, & Liu, 2022a) sample-specific anchors.

As shown in Fig. 7(b), fixed general face anchors achieved the best average AUC of 90.9%, outperforming both CoOp-style anchors and CoCoOp-style anchors. CoOp-style learnable anchors still improved over disabling SCR, indicating that learnable prompts can provide certain semantic regularization. However, they were less stable than fixed general face anchors under cross-dataset evaluation. For example, CoOp-style anchors slightly improved performance on CDF-v1 and DFD, but showed clear degradation on CDF-v2, DFDC, and DFDCP. This suggests that learnable anchors may absorb source-domain biases from FF++ training data, which weakens their cross-dataset robustness. CoCoOp-style sample-specific anchors performed even worse, with an average AUC lower than that of the variant without SCR. Although sample-specific prompts introduce stronger flexibility, they also make the semantic reference vary with each input image. This is less aligned with the goal of SCR, which aims to preserve a stable semantic relation between the adapted visual representation and the original frozen CLIP space. These results indicate that SCR benefits more from stable and domain-agnostic semantic anchors than from highly adaptive prompt parameterization. Therefore, we adopt fixed general face prompts as the default anchors for their simplicity, stability, and stronger cross-dataset generalization.

4.8 Model Complexity and Efficiency

Table 6 compares the inference efficiency of representative detectors. We report inference parameters, trainable parameters, FLOPs, and FPS. Here, inference parameters denote the parameters involved in the inference forward pass, trainable parameters denote the parameters updated during training, FLOPs are measured per frame, and FPS denotes frame-level inference throughput at batch size 1. Xception and UCF are included as CNN-based baselines, while Effort serves as the direct CLIP-based baseline of SemFusion.

As expected, CLIP-based detectors require more FLOPs than CNN-based models due to the large ViT-L/14 backbone. SemFusion increased the inference parameters only slightly from 304.17M to 305.52M and the FLOPs from 81.18G to 81.46G. The inference speed decreased mildly from 93.95 FPS to 89.29 FPS. This indicates that the proposed local evidence branch introduces limited additional inference overhead. With only 1.55M trainable parameters, SemFusion remains parameter-efficient while achieving strong cross-dataset and cross-method generalization.

5 Conclusion

In this paper, we have proposed SemFusion for generalizable FFD. SemFusion is designed to maintain semantic consistency during CLIP visual adaptation and strengthen the modeling of sparse localized forgery evidence. Instead of using textual prompts as direct real/fake classifiers, SemFusion has introduced SCR to regularize the adapted visual representation with CLIP text anchors as semantic references. To capture localized artifacts, SemFusion introduces MPE, which mines suspicious patch responses from selected visual layers under image-level supervision. In inference, the local fake-evidence logit is fused with the global CLS prediction, enabling local evidence to complement the forged face decision. Extensive experiments have shown that SemFusion achieved strong generalization performance, with an average frame-level AUC of 0.909 on five cross-dataset benchmarks and an average video-level AUC of 0.974 on six unseen forgery methods. Ablation studies, robustness evaluation, visualization, and efficiency analysis further confirm the complementary roles of semantic regularization and local evidence learning. Future work will explore temporal forgery reasoning and improve robustness against severe compression and emerging generative manipulation techniques.

6 Supplementary Information

Acknowledgment. This work was supported in part by the General Research Fund of Research Grants Council under grants: 12201323, 12202025, and 12202924.

Author Contributions. All authors contributed to the research conception and design. Framework implementation and analysis were performed by Mingjie Zhao. The first draft of the manuscript was written by Mingjie Zhao and Guilin Pang. Yiu-ming Cheung supervised the research and provided overarching guidance. All authors critically reviewed and refined previous versions of the manuscript, and all authors read and approved the final manuscript.

Data Availability. This work uses publicly available or research-accessible face forgery datasets following their original licenses and terms of use. The source training dataset, FaceForensics++ (FF++), together with the Google/Jigsaw DeepFakeDetection (DFD) dataset, is available through the FaceForensics++ project at https://github.com/ondyari/faceforensics. Celeb-DF-v1 and Celeb-DF-v2 are available at https://cse.buffalo.edu/~siweilyu/celeb-deepfakeforensics.html. The DeepFake Detection Challenge (DFDC) dataset and the DFDC Preview (DFDCP) dataset are available at https://ai.meta.com/datasets/dfdc/. The cross-method evaluation involving UniFace, BlendFace, MobileSwap, E4S, FaceDancer, and FSGAN follows the publicly released DeepfakeBench/DF40 benchmark resources at https://github.com/YZY-stack/DF40. All source code and evaluation scripts will be made publicly available upon acceptance of this paper.

Declarations

Conflicts of Interest. The authors declare that they have no conflict of interest.

Ethical Approval. This study used publicly available datasets and previously collected data and did not collect new data involving human or animal subjects, so ethical approval or informed consent was not required.

Consent for Publication. All authors agreed with the content and gave explicit consent for the submission and publication of this work.

Consent to Participate. Not applicable.

References

Ba, Z., Liu, Q., Liu, Z., Wu, S., Lin, F., Lu, L., Ren, K. (2024). Exposing the deception: Uncovering more forgery clues for deepfake detection. *Proceedings of the aaai conference on artificial intelligence* (Vol. 38, pp. 719–728).

Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., Yang, X. (2022). End-to-end reconstruction-classification learning for face forgery detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4113–4122).

Chen, L., Zhang, Y., Song, Y., Liu, L., Wang, J. (2022). Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 18710–18719).

Chen, Y., Yan, Z., Cheng, G., Zhao, K., Lyu, S., Wu, B. (2026). X2-dfd: A framework for explainable and extendable deepfake detection. *Advances in Neural Information Processing Systems*, *38*, 83792–83839,

Cheng, J., Yan, Z., Zhang, Y., Luo, Y., Wang, Z., Li, C. (2024). Can we leave deepfake data behind in training deepfake detector? *Advances in Neural Information Processing Systems*, *37*, 21979–21998,

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1251–1258).

Cui, X., Li, Y., Luo, A., Zhou, J., Dong, J. (2025). Forensics adapter: Adapting clip for generalizable face forgery detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 19207–19217).

Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K. (2020). On the detection of digital face manipulation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5781–5790).

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C. (2020). *The deepfake detection challenge (dfdc) dataset.*

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C. (2019). *The deepfake detection challenge (dfdc) preview dataset.*

Fei, H., Jia, Z., Huang, C., Zhang, J., Zhou, J. (2026). *Exploring specular reflection inconsistency for generalizable face forgery detection.*

Fu, X., Yan, Z., Yao, T., Chen, S., Li, X. (2025). Exploring unbiased deepfake detection via token-level shuffling and mixing. *Proceedings of the aaai conference on artificial intelligence* (Vol. 39, pp. 3040–3048).

Guo, X., Song, X., Zhang, Y., Liu, X., Liu, X. (2025). Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. *Proceedings of the computer vision and pattern recognition conference* (pp. 105–116).

Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., Ye, D. (2023). Implicit identity driven deepfake face swapping detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4490–4499).

Karras, T., Laine, S., Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4401–4410).

Khan, S.A., & Dang-Nguyen, D.-T. (2024). Clipping the deception: Adapting vision-language models for universal deepfake detection. *Proceedings of the international conference on multimedia retrieval* (pp. 1006–1015).

Kingma, D.P. (2014). *Adam: A method for stochastic optimization.*

Lai, Y., Yu, Z., Wang, J., Shen, L., Xu, Y., Cao, X. (2026). Agent4faceforgery: Multi-agent llm framework for realistic face forgery detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 14073–14083).

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B. (2020). Face x-ray for more general face forgery detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5001–5010).

Li, Q., Wang, W., Xu, C., Sun, Z., Yang, M.-H. (2024). Learning disentangled representation for one-shot progressive face swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 8348-8364,

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3207–3216).

Lin, K., Lin, Y., Li, W., Yao, T., Li, B. (2025). Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. *Proceedings of the aaai conference on artificial intelligence* (Vol. 39, pp. 5262–5270).

Lin, Y., Song, W., Li, B., Li, Y., Ni, J., Chen, H., Li, Q. (2024). Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. *European conference on computer vision* (pp. 104–122).

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., ... Yu, N. (2021). Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 772–781).

Luo, A., Kong, C., Huang, J., Hu, Y., Kang, X., Kot, A.C. (2024). Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, *19*, 1168–1182,

Luo, Y., Zhang, Y., Yan, J., Liu, W. (2021). Generalizing face forgery detection with high-frequency features. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 16317–16326).

Ma, L., Yan, Z., Xu, J., Chen, Y., Guo, Q., Bi, Z., ... others (2025). From specificity to generality: Revisiting generalizable artifacts in detecting face deepfakes. *Advances in Neural Information Processing Systems*, *38*, 69306–69344,

Nguyen, D., Mejri, N., Singh, I.P., Kuleshova, P., Astrid, M., Kacem, A., ... Aouada, D. (2024). Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 17395–17405).

Ni, Y., Meng, D., Yu, C., Quan, C., Ren, D., Zhao, Y. (2022). Core: Consistent representation learning for face forgery detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12–21).

Nirkin, Y., Wolf, L., Keller, Y., Hassner, T. (2022). Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6111-6121,

Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. *European conference on computer vision* (pp. 86–103).

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning* (pp. 8748–8763).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10684–10695).

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1–11).

Shao, R., Wu, T., Nie, L., Liu, Z. (2025). Deepfake-adapter: Dual-level adapter for deepfake detection. *International Journal of Computer Vision*, *133*(6), 3613–3628,

Shiohara, K., & Yamasaki, T. (2022). Detecting deepfakes with self-blended images. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 18720–18729).

Sun, K., Chen, S., Yao, T., Zhou, Z., Ji, J., Sun, X., ... Ji, R. (2025). Towards general visual-linguistic face forgery detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 19576–19586).

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Niessner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2387–2395).

Tian, J., Chen, P., Yu, C., Fu, X., Wang, X., Dai, J., Han, J. (2024). Learning to discover forgery cues for face forgery detection. *IEEE Transactions on Information Forensics and Security*, *19*, 3814-3828,

Wang, C., He, Z., Hu, X., Guan, W., Wang, W., Fu, Z. (2026). Map-mamba: Multi-artifacts perception mamba for generalizable face forgery detection. *IEEE Transactions on Information Forensics and Security*, ,

Yan, Z., Luo, Y., Lyu, S., Liu, Q., Wu, B. (2024). Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8984–8994).

Yan, Z., Wang, J., Jin, P., Zhang, K.-Y., Liu, C., Chen, S., ... Yuan, L. (2025). Orthogonal subspace decomposition for generalizable ai-generated image detection. *International conference on machine learning* (pp. 70268–70288).

Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., ... others (2024). Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, *37*, 29387–29434,

Yan, Z., Zhang, Y., Fan, Y., Wu, B. (2023). Ucf: Uncovering common features for generalizable deepfake detection. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 22412–22423).

Yan, Z., Zhang, Y., Yuan, X., Lyu, S., Wu, B. (2023). Deepfakebench: A comprehensive benchmark of deepfake detection. *Advances in Neural Information Processing Systems*, *36*, 4534–4565,

Yan, Z., Zhao, Y., Chen, S., Guo, M., Fu, X., Yao, T., ... Yuan, L. (2025). Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. *Proceedings of the computer vision and pattern recognition conference* (pp. 12615–12625).

Zhang, Y., Wang, T., Yu, Z., Gao, Z., Shen, L., Chen, S. (2025). Mfclip: Multi-modal fine-grained clip for generalizable diffusion face forgery detection. *IEEE Transactions on Information Forensics and Security*, ,

Zhou, K., Yang, J., Loy, C.C., Liu, Z. (2022a). Conditional prompt learning for vision-language models. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 16816–16825).

Zhou, K., Yang, J., Loy, C.C., Liu, Z. (2022b). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2337–2348,