ResProto-FD: Visual-Language Residual Prototype Sets for Generalized Face Forgery Detection

Jiuyao Jing[1], **Yu Zheng**[1], **Chunlei Peng**[1]*

[1]School of Cyber Engineering, Xidian University, Xi'an, China
25151110112@stu.xidian.edu.cn, yzheng@xidian.edu.cn, clpeng@xidian.edu.cn

Abstract

With the rapid development of generative models, such as generative adversarial networks and diffusion models, the task of face forgery detection has emerged, aiming to identify forged faces in real-world scenarios. A key challenge for current face forgery detection models is improving generalization to unknown forgeries. To address this, we propose ResProto-FD, a framework that constructs residual prototype sets to capture diverse forgery cues and discriminative differences from real faces. Our novel perspective collects prototypes from the most informative residual features generated during training, enabling better representation of various forgery traces and real-vs-fake distinctions. First, we introduce a Visual-Language Residual Learning (VLRL) module based on the CLIP model. This module constructs residual features between image and text embeddings to capture inconsistencies between visual features and associated textual semantics. In doing so, it guides the model to attend to subtle visual forgery clues and enhances the discriminative power of image representations. Furthermore, we design a Gradient-aware Residual Prototypes (GRP) mechanism— a dynamic collection strategy that selectively stores uncertain residual features based on gradient signals to build the prototype sets. This enhances the model's ability to generalize to unknown forgery types. Extensive experiments across various datasets and forgery methods demonstrate that ResProto-FD significantly improves generalization performance and consistently outperforms state-of-the-art methods.

Introduction

With the rapid advancement of generative models such as generative adversarial networks and diffusion models, face forgery techniques have become increasingly realistic, posing serious threats to privacy, identity security, and public trust, particularly in authentication scenarios.

In response to these concerns, face forgery detection has emerged as a critical research task. It is typically formulated as a binary classification problem that aiming to distinguish between authentic and manipulated facial inputs. While existing methods have shown promising results within specific datasets, their generalization to unseen forgery types remains limited (Yan et al. 2023b, 2024b). As forgery techniques evolve rapidly, improving the generalization capability of detection models has become a pressing challenge. This issue parallels the broader out-of-distribution (OOD) detection problem (Shu et al. 2023), which evaluates a model's robustness when encountering novel data distributions.

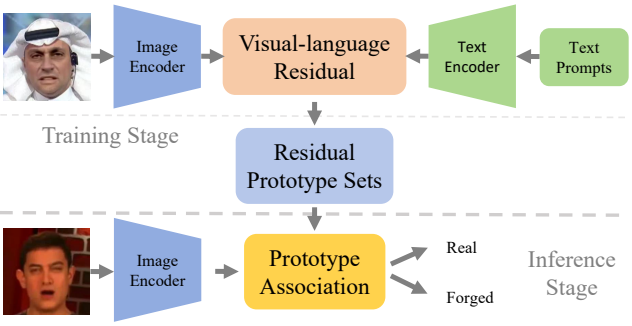

Figure 1: We construct the prototype sets using the residual features obtained from the visual-language residual calculations shown during training in the top half. During inference, we use prototype association to evaluate the match between test image features and prototypes, thereby distinguishing real faces from forged ones.

The recent development of multimodal vision-language models (such as CLIP (Radford et al. 2021)) has introduced new opportunities for face forgery detection. By jointly modeling visual and textual semantics, these models can capture high-level abstractions of forgery cues, potentially enhancing both robustness and generalization. Consequently, recent studies have begun to explore multimodal alignment not only to improve detection performance but also to enable more interpretable forgery analysis. However, existing CLIP-based methods primarily focus on the quality and diversity of text prompts, neglecting subtle image artifacts in semantically uncovered areas of the text prompts, resulting in limited generalization ability for detecting unseen forgery types. Furthermore, current approaches often emphasize the model's internal ability to represent broad forgery features, while exploration of how to further represent unknown forgery methods based on existing forgery im-

*Corresponding author.

age features (e.g., constructing prototype sets) remains very limited.

To bridge this gap, we propose ResProto-FD (Visual-Language Residual Prototype Sets) for generalizable face forgery detection. ResProto-FD dynamically constructs prototype sets using residual features obtained during training. As illustrated in Fig. 1, the framework operates in two stages: training and inference. In the training stage, visual-linguistic residuals between image and text features are computed and aligned with category-specific text embeddings, compelling the model to attend to forgery cues beyond textual semantics and enhancing the discriminability of image representations. Concurrently, residual features are analyzed to identify samples with high misclassification risk, enabling the formation of real and forged residual prototype sets. During inference, the cosine similarity between an input image feature and prototypes in each set is used for classification. This design effectively generalizes to detect forgeries produced by previously unseen manipulation techniques.

Our main contributions are as follows:

- We propose a novel face forgery detection framework called ResProto-FD, which aims to leverage existing image representations to further enhance the generalization ability of forgery detection. To this end, we introduce Visual-Language Residual Learning (VLRL) to guide the model toward forgery clues in regions of the image representation not covered by the text, enhancing the discriminative ability of image representations for forgery clues.
- We introduce Gradient-aware Residual Prototypes (GRP), a dynamic collection mechanism designed to extract prototypes that improve generalization by leveraging residual features generated during training. GRP uses gradient-based feature selection and clustering to retain prototypes with uncertain categories. Outdated prototypes are removed through a score decay strategy, maintaining high-quality prototype sets.
- We conduct extensive cross-dataset and cross-forgery evaluations to validate the generalization of the proposed ResProto-FD framework. Experimental results demonstrate that our method achieves competitive performance compared to state-of-the-art approaches.

Related Work

Face Forgery Methods

With the rapid advancement of generative models, facial forgery techniques, commonly referred to as deepfakes, have witnessed remarkable improvements in photorealism and controllability. These techniques can be broadly categorized into four types: face swapping (FS), facial reenactment (FR), entire face synthesis (EFS), and face editing (FE). FS replaces the facial identity while preserving pose and lighting; FR transfers expressions across identities; EFS generates entirely synthetic faces from scratch; and FE modifies facial attributes such as age or gender while maintaining identity consistency.

Such forgery techniques are predominantly driven by generative adversarial networks, such as StyleGAN (Karras, Laine, and Aila 2019), and more recently by diffusion models like Stable Diffusion (Rombach et al. 2022), which are capable of learning high-fidelity facial priors. To support research in this field, recent benchmarks such as DF40 have emerged, offering diverse and high-quality deepfake samples spanning 40 manipulation methods. Notably, DF40 incorporates 40 challenging forgery methods, including six representative FS methods: SimSwap, InSwap, UniFace, e4s, FaceDancer, and FSGAN, each of which has its own strengths and weaknesses in identity preservation, temporal coherence, and artifact suppression. These diverse characteristics pose new challenges for robust forgery detection.

Face Forgery Detection

Face forgery detection methods can be classified by input modality into image-level and video-level approaches. Image-level methods focus on detecting manipulation artifacts from individual frames. Early techniques relied heavily on handcrafted features (Li, Chang, and Lyu 2018; Yang, Li, and Lyu 2019), whereas more recent approaches leverage spatial-frequency representations (Zhou et al. 2024; Kashiani, Talemi, and Afghah 2025), reconstruction-based cues (Cao et al. 2022; Sun et al. 2024), data-augmentation based (Yan et al. 2024a; Lin et al. 2024b), or disentangled representations (Yan et al. 2023a; Cheng et al. 2025) to improve generalization under distribution shifts. In contrast, video-level methods (Zheng et al. 2021; Larue et al. 2023; Zhang et al. 2024) utilize inter-frame temporal consistency to detect, targeting subtle motion inconsistencies or unnatural transitions in expressions and head movements. These methods typically emphasize the model's ability to generalize and represent image features within the model itself, while there is relatively little research on how to further improve generalization based on existing image features (e.g., by constructing prototype sets).

CLIP-Based Methods

The success of vision-language models like CLIP (Radford et al. 2021) has spurred a new wave of research in face forgery detection. By embedding images and text into a shared semantic space, these models enable cross-modal reasoning and improved generalization. Several works have explored leveraging this capability: RepDFD (Lin et al. 2025) enhance CLIP features using learnable visual perturbations and adaptive prompts derived from facial embeddings; C2P-CLIP (Tan et al. 2025) incorporates category-aware prompts into the text encoder for better visual-textual alignment; VLFFD (Sun et al. 2025) employs a forgery-oriented text generator to guide a multimodal LLM in producing fine-grained descriptions of manipulated regions. Meanwhile, studies such as (Shi et al. 2025) have investigated explainable CLIP-based detection frameworks. Although these methods are very effective, they do not fully explore the image features within the CLIP model, and still focus on semantic coverage of text descriptions and image-text alignment.

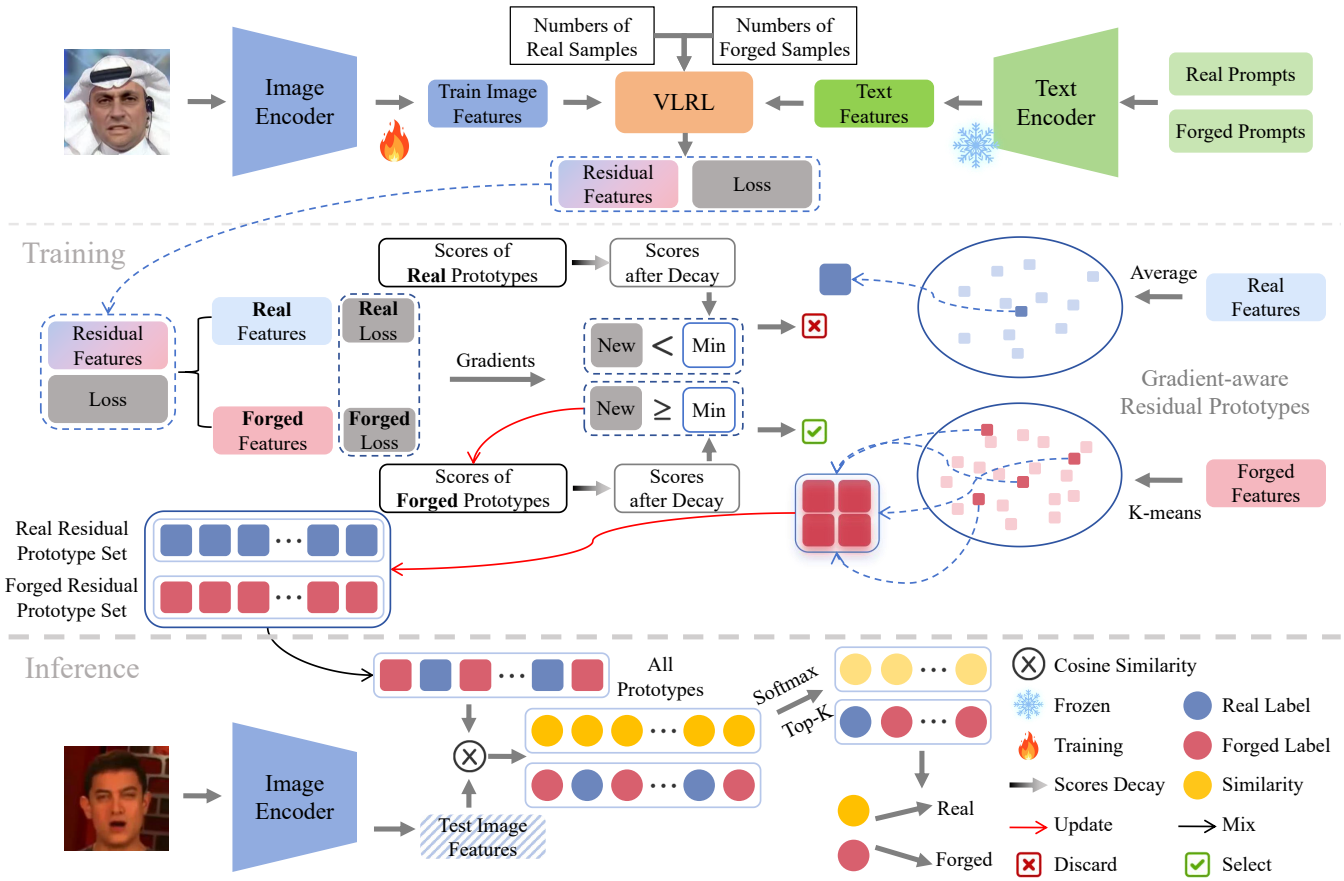

Figure 2: The overall structure of ResProto-FD. The upper section consists of two modules, representing the model's training stage. The upper region of the training stage represents the CLIP model training process, including the VLRL, while the lower region shows the GRP workflow. The lower section illustrates the ResProto-FD inference stage.

Methodology

Training Stage

The overall architecture of the proposed **ResProto-FD** model is shown in Fig. 2 . The upper part of the figure describes the training stage. Let E_T and E_I denote the text encoder and image encoder of the CLIP model, respectively. The text prompts $P^{(c)}$ are shown in Tab. 1, where $c \in 0, 1$ represents the class label (0 for real and 1 for forged). Given a training face image $x_{train}^{(c)}$, the corresponding image feature is extracted as $f_{train}^{(c)} = E_I(x_{train}^{(c)})$. After being encoded by the text encoder, the corresponding text feature is calculated as $f_{text}^{(c)} = E_T(P^{(c)})$ (for multiple text prompts within a single class, the average feature is taken). During training, the text encoder E_T is kept frozen and only the image encoder E_I is updated.

Visual-Language Residual Learning (VLRL) The internal structure of the VLRL is illustrated in Fig. 3. To capture discriminative clues that are not semantically aligned with the textual description, we compute a visual-language residual feature $f_{res}^{(c)}$ (hereinafter referred to as residual features) by subtracting a scaled version of the class-specific text feature $f_{text}^{(c)}$ from the image feature:

$$f_{res}^{(c)} = f_{train}^{(c)} - \lambda f_{text}^{(c)} \tag{1}$$

Here, λ is a hyperparameter that controls the semantic coverage of the text embedded in the image. The residual feature $f_{res}^{(c)}$ is designed to enhance the discriminative ability of image representations for forgery clues by emphasizing the discrepancy between visual and textual modalities.

Next, we compute the similarity between the residual feature and the class-specific text feature to obtain the final classification logits $z_{res}^{(c)}$:

$$z_{res}^{(c)} = sim(f_{res}^{(c)}, f_{text}^{(c)}) \cdot \tau \tag{2}$$

where $sim(\cdot)$ denotes a similarity function, such as cosine similarity, and τ is a learnable temperature parameter that adjusts the sharpness of the logit distribution. This formulation encourages the model to attend to semantic misalignments that may be indicative of facial manipulations, thereby improving its discriminative capability and generalization performance.

To further enhance the discriminative ability of image representation for forgery cues, we introduce a regularization term for residual learning (L_{RL}), which integrates class

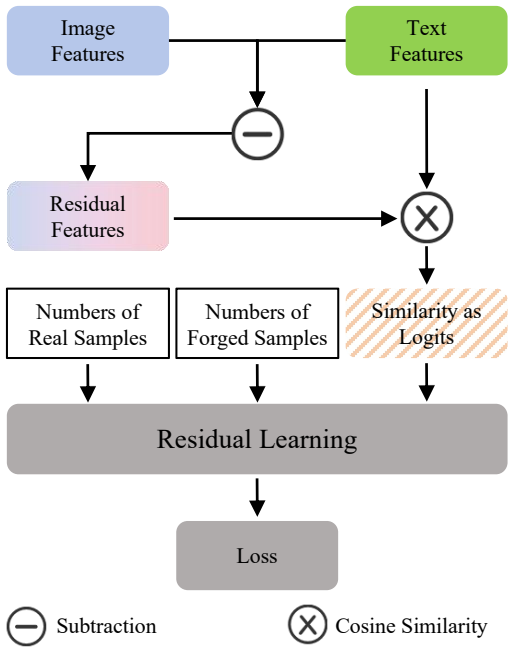

Figure 3: Schematic diagram of the VLRL.

prior correction and mutual information (MI) regularization (Hong et al. 2021). This design facilitates more generalized representation learning.

Let $p^{(c)} = \frac{N^{(c)}}{N^{(0)}+N^{(1)}}$ denote the empirical prior of class $c \in 0, 1$, where $N^{(c)}$ is the number of training samples of class c, and the target prior is assumed uniform $b^{(c)} = \frac{1}{2}$. To correct for imbalance, we compute a class-aware cross-entropy loss with log-prior adjustment:

$$L_{ce}^{(c)} = CE(z_{res}^{(c)} + \log(p^{(c)}), c) \tag{3}$$

where $z_{res}^{(c)}$ is the logits calculated by the cosine similarity between the residual features and the text features.

To promote class-discriminative residual learning, an MI-based regularization term is used. The positive and negative statistics are defined as:

$$z_{pos}^{(c)} = \frac{1}{B^{(c)} + \epsilon} \sum_{i=1}^{B^{(c)}} z_{res-i}^{(c)}, \tag{4}$$

$$z_{neg}^{(c)} = \log\left(\frac{1}{B}\sum_{i=1}^{B} \frac{b^{(c)}}{p^{(c)}} e^{z_{res-i}^{(c)}}\right) \tag{5}$$

where B represents the total number of all samples in a batch, and $B^{(c)}$ represents the total number of samples of category c in the batch, and ϵ represents the minimum value to prevent the denominator from being zero (1×10^{-9} is taken in the experiment). And the regularization loss is given by:

$$L_{reg}^{(c)} = -\left(z_{pos}^{(c)} - z_{neg}^{(c)} - \omega_1 (z_{neg}^{(c)})^2\right) \tag{6}$$

The final objective for the VLRL is:

$$L_{RL}^{(c)} = L_{ce}^{(c)} + \omega_2 \cdot L_{reg}^{(c)} \tag{7}$$

where $\omega 1$ and ω_2 are balancing hyperparameters.

Gradient-aware Residual Prototypes (GRP) The proposed GRP module is illustrated in the middle part of Fig. 2. To enhance the generalization capability of the model, particularly when encountering previously unseen forgery types, we introduce a dynamic feature selection and retention mechanism based on gradient sensitivity. GRP selectively retains the most informative residual features produced by the VLRL for each class, forming a prototype set that serves as a reference during inference.

Specifically, for each class $c \in 0, 1$ (where 0 indicates real and 1 indicates forged), we maintain a fixed-size prototype set $pts^{(c)}$ that can store at most $M^{(c)}$ prototypes and a prototype score set $g_{pts}^{(c)}$ of the same size. During training, for each incoming mini-batch, we compute the ℓ_2-norm of the gradient of the class-specific loss with respect to the corresponding residual feature as the prototype score $g^{(c)}$ for class c:

$$g^{(c)} = \left\| \frac{\partial L_{RL}^{(c)}}{\partial f_{res}^{(c)}} \right\|_2 \tag{8}$$

For real samples ($c = 0$), the candidate feature is obtained by computing the average of all residual features from real samples within the current batch. This averaged feature is then associated with a single protptype score $g^{(0)}$.

For forged samples ($c = 1$), due to their greater intra-class diversity, arising from different spoofing methods or manipulation styles, we apply K-means clustering on the residual features of the batch. The centers of the obtained sub-clusters are extracted as candidate prototypes, and the prototype score of each candidate prototype is $g^{(1)}$.

When the prototype set $pts^{(c)}$ has not reached its capacity, protptypes are inserted directly along with their associated scores. Once filled, we apply a score decay strategy that incorporates prototype dwell time to avoid retaining outdated or less informative prototypes. Specifically, let the protptype score of the new batch corresponding to the category be $g_{new}^{(c)}$. If the following conditions are met, the stored prototype $f_{pts-j}^{(c)}$ will be replaced by a new candidate prototype with protptype score $g_{new}^{(c)}$:

$$g_{new}^{(c)} \geq \min\left(g_{pts-j}^{(c)} \cdot \gamma^{t_{pts-j}^{(c)}}\right) \tag{9}$$

where $\gamma \in (0, 1)$ is a hyperparameter that controls the decay rate.

This replacement criterion ensures that the prototype set prioritizes retaining recent and highly informative residual features, while discarding outdated or low-confidence entries, even if they previously had strong gradient contributions. Therefore, GRP can alleviate the feature drift problem and maintain a dynamic prototype set for each category.

Inference Stage

During inference as shown in the lower part of Fig. 2, given a test image x_{test}, the trained image encoder E_I is used to extract its feature representation $f_{test} = E_I(x_{test})$.

The CLIP-based forged sample confidence score $Score_{clip}$ is calculated as the similarity $z_{res}^{(1)}$ between the

Class	Prompts
Real	This is an example of a real face This is a real face This is how a real face looks like a photo of a real face
Forged	This is an example of a forged face This is a synthetic face This is how a manipulated face looks like a photo of an artificial face

Table 1: Overview of real and forged text prompts.

residual features of the forged category and the text features of the corresponding category (See Eq. 2 for details).

After introducing GRP, we first mix all the features of the prototype sets of different categories together and record them as f_{GRP}, and then calculate the cosine similarity between the test image feature f_{test} and each prototype f_{GRP-i}:

$$sim_i = sim(f_{GRP-i}, f_{test}) \tag{10}$$

Next, we select the K highest similarity sim_K. These similarity scores are then normalized using the softmax function:

$$z_{GRP-K} = Softmax(sim_K) \tag{11}$$

At this point, find the category c of the corresponding prototype according to z_{GRP-K} and classify it according to the category to obtain $z^{(c)}_{GRP-K}$. The forged sample confidence score is calculated by adding the normalized similarity of the forged prototypes ($c = 1$):

$$z^{(1)}_{GRP} = \sum_{i=1}^{K^{(1)}} z^{(1)}_{GRP-i} \tag{12}$$

where $K^{(1)}$ is the number of forged class among the K retrieved prototypes.

Finally, the score $z^{(1)}_{GRP}$ is used as the forgery confidence score $Score_{GRP}$ derived from GRP.

Experiments

Datasets

To evaluate the generalization ability of the proposed model, we conduct experiments on several widely used face forgery detection datasets: FaceForensics++ (FF++) (Rossler et al. 2019), CelebDF-v2 (CDF) (Li et al. 2020b), DeepfakeDetection (DFD) (Research and Jigsaw 2019), Deepfake Detection Challenge (DFDC) (Dolhansky et al. 2019). All datasets are in the form of preprocessed frames (256×256) provided by previous studies (Yan et al. 2023b). In addition, we use the recently released DF40 dataset, which includes more forgery methods such as uniface, e4s, facedancer, fsgan, inswap and simswap. Following the experimental setup of prior works (Yan et al. 2023b; Chen et al. 2024), we train on the c23 compressed version of the FF++ dataset and test on the remaining datasets.

Methods	CDF	DFD	DFDC	Avg.
Xception*	73.7	81.6	70.8	75.4
EfficientB4*	74.9	81.5	69.6	75.3
Face X-ray*	67.9	76.6	63.3	69.3
F3Net*	73.5	79.8	70.2	74.5
FFD*	74.4	80.2	70.3	75.0
SPSL*	76.5	81.2	70.4	76.0
SRM*	75.5	81.2	70.0	75.6
CORE*	74.3	80.2	70.5	75.0
RECCE*	73.2	81.2	71.3	75.2
UCF*	75.3	80.7	71.9	76.0
FoCus	82.8	-	72.8	-
PFGDD	74.4	84.8	61.5	73.6
LSDA	83.0	88.0	73.6	81.5
DiffusionFake	80.5	90.4	-	-
RepDFD	80.0	-	77.3	-
Freqdebias	83.6	86.8	74.1	81.5
Ours	**83.8**	**91.4**	**79.5**	**84.9**

Table 2: Cross-dataset evaluation using the frame-level AUC metric. The symbol * indicates that the results are cited from (Yan et al. 2024a), and other results are cited from the original paper. All models are trained on the FF++ (c23) dataset. The best results are shown in bold.

Implementation Details

We use ViT-B/16 as the image encoder of the CLIP base model and fine-tune only the image encoder while freezing all parameters of the text encoder. The hyperparameter λ is set to 0.3. The maximum number of memory entries per category in the GRP is 64, and the decay factor γ is set to 0.99, Top-K is set to 64. For residual features of category 1 within each batch, we apply K-means clustering to obtain 4 subcenters. Following (Hong et al. 2021), both ω_1 and ω_2 are set to 0.1. We adopt the Adam optimizer with a learning rate of 5×10^{-7} and a weight decay of 1×10^{-6}. Experiments are conducted using CUDA 12.4.1 and PyTorch 2.5.1 on two RTX 3090 GPUs. During the training stage, we used a simple data augmentation method: (1) resize to 300×300 and then randomly crop to 224×224, scale=(0.8, 1); (2) apply Gaussian Blur, kernel size=(3, 3), sigma=(0.1, 1.5); (3) apply Random Erasing, scale=(0.02, 0.2), ratio=(0.3, 3.3). Inspired by (Srivatsan, Naseer, and Nandakumar 2023), we use text prompts in Tab. 1.

Generalization Evaluation

To demonstrate the generalization ability of our proposed model, we compare the proposed ResProto-FD framework with previous state-of-the-art models on the CDF, DFD, and DFDC datasets. These include: Xception (Rossler et al. 2019), EfficientB4 (Tan and Le 2019), Face X-ray (Li et al. 2020a), F3Net (Qian et al. 2020), FFD (Dang et al. 2020), SPSL (Liu et al. 2021), SRM (Luo et al. 2021), CORE (Ni et al. 2022), RECCE (Cao et al. 2022), UCF (Yan et al. 2023a), FoCus (Tian et al. 2024), PFGDD (Lin et al. 2024a), LSDA (Yan et al. 2024a), DiffusionFake (Sun et al. 2024),

Methods	CDF	DFD	DFDC	Avg.
Xception†	81.6	89.6	73.2	81.5
PCL+I2G†	90.0	-	67.5	-
LipForensics	82.4	-	73.5	-
FTCN	86.9	-	-	-
UIA-ViT	82.4	94.7	-	-
SBI†	90.6	87.7	75.2	84.5
SLADD†	79.7	-	77.2	-
CORE†	80.9	88.2	72.1	80.4
DCL†	88.2	92.1	75.0	85.1
UCF†	83.7	86.7	77.0	82.5
SeeABLE	87.3	-	75.9	-
NACO	89.5	-	76.7	-
LSDA	89.8	95.6	73.5	86.3
RepDFD	89.9	-	81.0	-
Ours	**91.5**	**95.8**	**81.9**	**87.5**

Table 3: Cross-dataset evaluation using the video-level AUC metric. The symbol †represents results are cited from (Ma et al. 2025), and other results are cited from the original paper. All models are trained on the FF++ (c23) dataset. The best results are shown in bold.

RepDFD (Lin et al. 2025), Freqdebias (Kashiani, Talemi, and Afghah 2025), PCL+I2G (Zhao et al. 2021), LipForensics (Haliassos et al. 2021), FTCN (Zheng et al. 2021), UIA-ViT (Zhuang et al. 2022), SBI (Shiohara and Yamasaki 2022), SLADD (Chen et al. 2022), DCL (Sun et al. 2022), SeeABLE (Larue et al. 2023), NACO (Zhang et al. 2024), IID (Huang et al. 2023), CDFA (Lin et al. 2024b), ProDet (Cheng et al. 2024) and X^2DFD (Chen et al. 2024).

Frame-level AUC comparison results are summarized in Tab. 2. It is evident that the proposed ResProto-FD framework consistently achieves superior performance across all datasets in the generalization experiments. The video-level AUC results, presented in Tab. 3, further confirm this trend, as our method consistently ranks first across all benchmarks. These findings strongly validate the effectiveness of the proposed prototype set in enhancing the generalization capability of face forgery detection models.

To further evaluate the generalization ability of ResProto-FD, we conducted experiments on the DF40 dataset, selecting four representative forgery methods: e4s, fsgan, inswap, and simswap. The corresponding results are presented in Tab. 4. As shown, our model surpasses all competing approaches in terms of AUC, clearly demonstrating the superior generalization of the proposed ResProto-FD framework when faced with previously unseen forgery methods.

Ablation Study

Modules Analysis We conducted ablation studies on the proposed VLRL and GRP modules across multiple datasets to assess their individual and combined contributions. As shown in Tab 5, the evaluation includes average frame-level results on the CDF, DFD, and DFDC datasets, as well as on the DF40 dataset across six forgery methods (Uniface, E4S, FaceDancer, FSGAN, InSwap, and SimSwap). All experiments adopt the same data augmentation strategy described in the **Implementation Details**.

Methods	e4s	fsgan	inswap	simswap	Avg.
RECCE	65.2	88.4	79.5	73.0	76.5
SBI	69.0	87.9	63.3	56.8	69.3
CORE	63.4	91.1	79.4	69.3	75.8
IID	71.0	86.4	74.4	64.0	74.0
UCF	69.2	88.1	76.8	64.9	74.8
LSDA	68.4	83.2	81.0	72.7	76.3
CDFA	67.4	84.8	72.0	76.1	75.1
ProDet	71.0	86.5	78.8	77.8	78.5
X^2-DFD	91.2	89.9	78.4	84.9	86.1
Ours	**94.8**	**94.9**	**89.3**	**85.7**	**91.2**

Table 4: Cross-dataset evaluation using the frame-level AUC metric on the DF40 dataset. All models were trained on the FF++ (c23) dataset. The best results are shown in bold.

V	G	CDF & DFD & DFDC			DF40		
		AUC	AP	EER	AUC	AP	EER
×	×	83.1	89.2	24.8	86.0	91.8	21.5
✓	×	83.2	88.6	24.5	88.9	93.4	18.6
×	✓	83.1	89.4	24.5	85.9	92.4	21.3
✓	✓	**84.9**	**90.6**	**23.5**	**90.5**	**94.9**	**16.8**

Table 5: Frame-level ablation studies under different modules. All experiments are trained on FF++ (c23) and tested on the average metrics under two generalization experiments. V refers to VLRL and G refers to GRP.

The results indicate that the VLRL module notably enhances the generalization capability of the CLIP-based detection model. Although the GRP module alone does not yield a substantial gain in AUC, it still improves the generalization performance in terms of AP and EER compared to the baseline, suggesting its effectiveness in refining the feature space. When both modules are applied jointly, the model achieves the best overall performance across all metrics and settings, demonstrating that prototypes derived from the residual features induced by VLRL provide more discriminative representations of forged behaviors than prototypes derived from original image features.

Components of VLRL Analysis Tab. 6 presents ablation results for different components of the VLRL module under the same settings as Tab. 5. When only the residual learning regularization term is applied, the residual features f_{res} are replaced by the original CLIP image features f_{train} for cosine similarity computation. Furthermore, when training using only residual features, the loss function is replaced by the original cross-entropy loss. The results in Tab. 6 show that compared to the full ResProto-FD, using only the regularization term for residual learning results in the largest drop. Adding the residual features f_{res} effectively mitigates

R	L	CDF & DFD & DFDC			DF40		
		AUC	AP	EER	AUC	AP	EER
×	×	83.1	**89.2**	24.8	86.0	91.8	21.5
✓	×	82.9	88.8	24.8	86.2	91.6	21.1
×	✓	79.0	84.3	27.9	81.4	88.1	26.4
✓	✓	**83.2**	88.6	**24.5**	**88.9**	**93.4**	**18.6**

Table 6: Frame-level ablation studies of different VLRL components excluding GRP. The experimental setup follows Tab 5. R refers to f_{res}, L refers to L_{RL}.

this drop. Furthermore, compared to the results in the first row, using only the residual features f_{res} and only the regularization term for residual learning L_{RL} results in lower performance than the baseline. Using the full VLRL, the results significantly improve over the baseline. These observations suggest that only when the residual features and the regularization term of residual learning are used in conjunction can the model learn the inter-class differences between real and fake faces more effectively.

In fact, residual features allow the model to focus more on forgery cues in regions of the image not covered by text semantics and, through data augmentation, to better leverage the diversity of image samples. Meanwhile, the L_{RL} term enforces mutual information–based alignment between residual and text features, increasing the sensitivity of visual representations to forgery artifacts.

Score Decay of GRP We report generalization results for different decay rates in GRP in Tab. 7. The experimental setup follows Tab. 5 and Tab. 6. The results show that when $\gamma = 1.0$, it means that GRP does not introduce a score decay mechanism at this time. The best overall performance is achieved at $\gamma = 0.99$, validating the effectiveness of introducing the decay mechanism. This can be attributed to feature drift that naturally occurs when constructing prototype sets during training. However, by comparing the experimental results of $\gamma = 0.90$ and $\gamma = 0.95$, we can find that even if there is feature drift, its negative impact is limited under the influence of the gradient-aware strategy. Specifically, the smaller the γ parameter, the more serious the feature drift phenomenon is, which in turn has a greater impact when comparing prototype scores. However, compared with the improvement of the module on the overall performance (Tab. 5), the feature drift phenomenon has little effect on the gradient-aware feature selection.

K Parameter Discussion Fig. 4 illustrates the effect of the K hyperparameter on inference performance, with AUC results reported for $K = 16, 32, 64, 128$. Notably, 64 is equivalent to the number of prototypes in the prototype set of each category, and 128 is equivalent to the total number of prototypes. From the results we can see that the parameter K is not the bigger the better, nor the smaller the better. This is consistent with our intuition. When K is small, the number of matching prototypes decreases, and for forged samples, the confidence in the forged class decreases. As K increases,

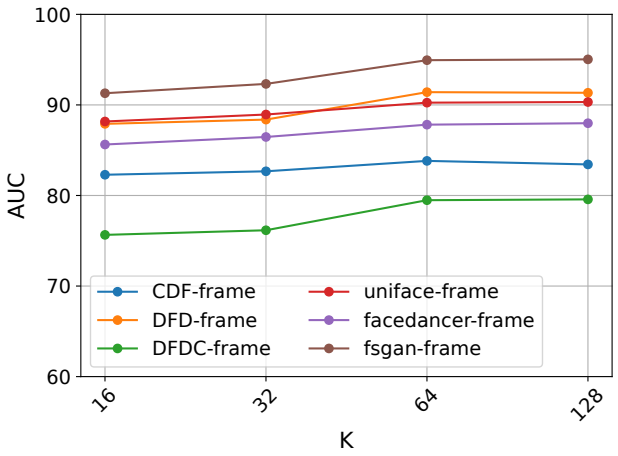

Figure 4: Statistics results under different K.

γ	CDF & DFD & DFDC			DF40		
	AUC	AP	EER	AUC	AP	EER
1.0	84.8	90.6	23.6	90.1	94.8	17.1
0.90	84.7	90.5	23.7	90.1	94.7	17.0
0.95	84.8	90.6	23.7	90.2	94.9	17.0
0.99	**84.9**	**90.6**	**23.5**	**90.5**	**94.9**	**16.8**

Table 7: Frame-level ablation studies of different decay rates γ in GRP. The experimental setup follows Tab 5.

the number of matching prototypes increases, and the confidence in the forged class changes as the number of matching prototypes of the true class increases. The watershed value is 64, meaning that ideally, the confidence reaches its highest value when the top 64 prototypes with the highest feature similarity to the test image are from the same class. In general, as shown in the Fig. 4, the AUC results stabilize at 64, so we choose 64 as the final K.

Conclusion

In order to fully explore the generalization of the CLIP model's image features, we proposed VLRL, which enables the model to mine important forgery clues in image areas not covered by text semantics through visual-language residual calculation. In addition, we introduced the GRP, combining gradient-aware with score decay and the prototype sets were robustly and dynamically constructed to store the most informative residual features, in order to further utilize the obtained generalized residual features in the inference stage. While our framework significantly improves generalization, it also significantly increases model training overhead. This is due to the additional backpropagation required to analyze gradients during training, as well as the increased training overhead of clustering or averaging features across a batch to obtain prototypes. This challenge will be a focus of our future research.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62276198, Grant U22A2035; in part by the Natural Science Basic Research Program of Shaanxi under Grant 2025JC-YBMS-696; in part by the Independent Research Topic Program of Shaanxi Key Laboratory of Intelligent Policing (Shaanxi Police College) under Item number SXZJ25ZZ02; in part by the innovation capability support plan in Shaanxi Province under Grant 2025ZC-KJXX-22; in part by Young Elite Scientists Sponsorship Program by CAST under Grant 2022QNRC001; in part by Open Research Project of Key Laboratory of Artificial Intelligence Ministry of Education under Grant AI202401; supported by the 111 Center (B16037) and in part supported by the Fundamental Research Funds for the Central Universities under Grant YJSJ25011.

References

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

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

Chen, Y.; Yan, Z.; Cheng, G.; Zhao, K.; Lyu, S.; and Wu, B. 2024. X2-dfd: A framework for explainable and extendable deepfake detection. *arXiv preprint arXiv:2410.06126*.

Cheng, J.; Yan, Z.; Zhang, Y.; Hao, L.; Ai, J.; Zou, Q.; Li, C.; and Wang, Z. 2025. Stacking brick by brick: Aligned feature isolation for incremental face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13927–13936.

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

Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; and Jain, A. K. 2020. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, 5781–5790.

Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; and Ferrer, C. C. 2019. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*.

Haliassos, A.; Vougioukas, K.; Petridis, S.; and Pantic, M. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5039–5049.

Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; and Chang, B. 2021. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6626–6636.

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

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kashiani, H.; Talemi, N. A.; and Afghah, F. 2025. Freqdebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8775–8785. IEEE.

Larue, N.; Vu, N.-S.; Struc, V.; Peer, P.; and Christophides, V. 2023. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21011–21021.

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

Li, Y.; Chang, M.-C.; and Lyu, S. 2018. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, 1–7. Ieee.

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

Lin, K.; Lin, Y.; Li, W.; Yao, T.; and Li, B. 2025. Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5262–5270.

Lin, L.; He, X.; Ju, Y.; Wang, X.; Ding, F.; and Hu, S. 2024a. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16815–16825.

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

Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; and Yu, N. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 772–781.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16317–16326.

Ma, L.; Yan, Z.; Chen, Y.; Xu, J.; Guo, Q.; Huang, H.; Liao, Y.; and Lin, H. 2025. From specificity to generality: Revisiting generalizable artifacts in detecting face deepfakes. *arXiv preprint arXiv:2504.04827*.

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

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Research, G.; and Jigsaw. 2019. Contributing Data to Deepfake Detection Research. Technical report, Google AI.

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

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Shi, Y.; Gao, Y.; Lai, Y.; Wang, H.; Feng, J.; He, L.; Wan, J.; Chen, C.; Yu, Z.; and Cao, X. 2025. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *Visual Intelligence*, 3(1): 9.

Shiohara, K.; and Yamasaki, T. 2022. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18720–18729.

Shu, Y.; Guo, X.; Wu, J.; Wang, X.; Wang, J.; and Long, M. 2023. Clipood: Generalizing clip to out-of-distributions. In *International conference on machine learning*, 31716–31731. PMLR.

Srivatsan, K.; Naseer, M.; and Nandakumar, K. 2023. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19685–19696.

Sun, K.; Chen, S.; Yao, T.; Liu, H.; Sun, X.; Ding, S.; and Ji, R. 2024. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems*, 37: 101474–101497.

Sun, K.; Chen, S.; Yao, T.; Zhou, Z.; Ji, J.; Sun, X.; Lin, C.-W.; and Ji, R. 2025. Towards general visual-linguistic face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19576–19586.

Sun, K.; Yao, T.; Chen, S.; Ding, S.; Li, J.; and Ji, R. 2022. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2316–2324.

Tan, C.; Tao, R.; Liu, H.; Gu, G.; Wu, B.; Zhao, Y.; and Wei, Y. 2025. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7184–7192.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

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

Yan, Z.; Luo, Y.; Lyu, S.; Liu, Q.; and Wu, B. 2024a. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8984–8994.

Yan, Z.; Yao, T.; Chen, S.; Zhao, Y.; Fu, X.; Zhu, J.; Luo, D.; Wang, C.; Ding, S.; Wu, Y.; et al. 2024b. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37: 29387–29434.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023a. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22412–22423.

Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; and Wu, B. 2023b. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In Oh, A.; Neumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 4534–4565. Curran Associates, Inc.

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 8261–8265. IEEE.

Zhang, D.; Xiao, Z.; Li, S.; Lin, F.; Li, J.; and Ge, S. 2024. Learning natural consistency representation for face forgery video detection. In *European Conference on Computer Vision*, 407–424. Springer.

Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; and Xia, W. 2021. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15023–15033.

Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; and Wen, F. 2021. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15044–15054.

Zhou, J.; Li, Y.; Wu, B.; Li, B.; Dong, J.; et al. 2024. Freqblender: Enhancing deepfake detection by blending frequency knowledge. *Advances in Neural Information Processing Systems*, 37: 44965–44988.

Zhuang, W.; Chu, Q.; Tan, Z.; Liu, Q.; Yuan, H.; Miao, C.; Luo, Z.; and Yu, N. 2022. UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European conference on computer vision*, 391–407. Springer.