SCORE: SEMANTIC-CONDITIONED ORTHOGONAL REFINEMENT FOR GENERALIZABLE AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalizable forgery detection requires forensic evidence that remains reliable under shifts in content, acquisition conditions, and generative mechanisms. A central challenge is that the diagnostic meaning of many forensic cues depends on the visual context in which they occur: local texture, boundary, and frequency responses can vary with identity, pose, expression, scene structure, and image quality. This motivates a controlled use of semantic context when forming and refining detection decisions. We propose **SCORE**, a **Semantic-Conditioned Orthogonal Refinement** framework that organizes detection into a semantic-conditioned primary stage and a conservative refinement stage. SCORE first uses the current visual state to parameterize a gated residual and forms an independently supervised primary prediction. It then constructs a sample-conditioned low-dimensional basis and applies soft orthogonal attenuation to obtain a structured refinement view that preserves part of the basis-aligned information. A conditional refinement head operates on this detached representation, while uncertainty derived from the detached primary prediction controls the strength of a bounded logit correction. SCORE achieves an average video AUC of 0.8984 across five cross-dataset targets, an average video AUC of 0.9723 over eight unseen manipulations in DF40, and 99.70% mAP and 95.48% mAcc across 19 UniversalFakeDetect (UFD) test sources. These results support structured semantic conditioning and controlled refinement as an effective approach to forgery detection under diverse distribution shifts.

1 INTRODUCTION

Rapid advances in generative models and face manipulation introduce shifts in generation mechanisms, content distributions, and image processing. Training-domain texture, frequency, and blending patterns may change substantially under unseen generators and target domains. Prior work has improved cross-domain robustness through frequency and high-frequency modeling, data augmentation, and representation disentanglement (Qian et al., 2020; Liu et al., 2021; Luo et al., 2021; Shiohara & Yamasaki, 2022; Yan et al., 2023; 2024a). Vision foundation models provide pretrained representations with broad semantic coverage (Radford et al., 2021), while EFFORT further improves knowledge preservation through constrained parameter adaptation (Yan et al., 2025). Generalization further depends on interpreting forensic cues under unseen content conditions.

A key challenge is that the diagnostic value of forensic cues often depends on the semantic context in which they arise. Oral textures vary with facial expression and mouth state, responses near facial boundaries depend on pose, hair, and background structure, and high-frequency patterns are affected by compression, noise, and image quality. Excessive semantic invariance can suppress informative cue–context interactions, whereas unrestricted semantic coupling can amplify domain-specific correlations associated with identity, scene, and capture conditions. We formulate this challenge as **controlled semantic conditioning**: organizing forensic evidence according to the semantic context of each sample while constraining how contextual information influences cross-domain prediction.

Based on this principle, we propose **Semantic-Conditioned Orthogonal Refinement** (**SCORE**), consisting of two successive modules. **Semantic-Conditioned Primary Prediction** (**SCPP**) organizes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Representative false-negative recoveries across four target domains. Each column shows a deterministically selected fake video and its paired Baseline and SCORE fake probabilities.

sample-dependent primary forensic evidence under the current context and produces an independently supervised prediction. **Instance-Conditioned Soft-Orthogonal Refinement** (**ISOR**) constructs a sample-conditioned refinement representation by softly attenuating basis-aligned components while retaining a controlled fraction of the projected information. It then uses primary uncertainty to modulate a bounded logit correction.

Figure 1 illustrates false negatives recovered by SCORE across subtle high-quality forgeries, capture-domain shift, in-the-wild distribution shift, and compound degradation. The broader paired analysis in Sec. 5.4 examines both corrected and newly introduced errors, together with domain-dependent score shifts.

We evaluate SCORE under three complementary generalization protocols: cross-dataset face forgery detection under joint domain shifts, cross-method evaluation on DF40 with unseen face manipulations, and UniversalFakeDetect (UFD) across diverse generation tasks and generative paradigms. SCORE achieves an average video AUC of 0.8984 across five cross-dataset targets, an average video AUC of 0.9723 over eight unseen manipulations in DF40, and 99.70% mAP and 95.48% mAcc across 19 UFD test sources.

Our main contributions are:

- We introduce **controlled semantic conditioning** as a perspective for generalizable forgery detection, characterizing how semantic context shapes forensic cue interpretation and cross-domain prediction.
- We propose **SCORE**, composed of SCPP and Instance-Conditioned Soft-Orthogonal Refinement, which couples context-conditioned primary prediction with uncertainty-aware bounded refinement.
- We evaluate SCORE across cross-dataset, cross-manipulation, and cross-generation protocols, and analyze its cross-domain behavior through paired score shifts and cross-source threshold transfer.

2 Related Work

Generalizable forgery detection. Prior work extracts forensic cues from face-warping artifacts (Li & Lyu, 2019), blending boundaries (Li et al., 2020a), frequency and phase information (Qian et al., 2020; Liu et al., 2021), and high-frequency residuals (Luo et al., 2021). SBI (Shiohara & Yamasaki, 2022) broadens training forgery patterns through self-blending of real images, while LSDA (Yan et al., 2024a) introduces intra-domain and inter-domain variations in latent forgery features. FreqDebias (Kashiani et al., 2025) combines forgery-mixup augmentation with local and global consistency regularization to mitigate dependence on specific frequency bands. Video-based methods additionally exploit temporal cues through mouth-motion modeling with lipreading pretraining (Haliassos et al., 2021), alternating freezing of spatial and temporal parameters (Wang et al., 2023), and thumbnail layouts that preserve spatiotemporal relationships across consecutive frames (Xu et al., 2023).

Disentangled and reconstruction-guided representations. UCF (Yan et al., 2023) disentangles representations into forgery-irrelevant, manipulation-specific, and common forgery features, using

the common features for detection. DFS-GDD (Ye et al., 2024) extracts common forgery semantics through semantic disentanglement, supported by an adaptive high-pass filtering module and two-stage training. RECCE (Cao et al., 2022) jointly optimizes reconstruction and classification, learning compact representations through real-image reconstruction and using reconstruction discrepancies to guide detection. DiffusionFake (Sun et al., 2024) injects detector features into a frozen Stable Diffusion model during training to reconstruct the corresponding source and target images, encouraging the detector to encode identity information associated with both.

Foundation-model adaptation for forensics. CLIP (Radford et al., 2021) learns transferable visual representations through large-scale image–text pretraining. UniFD (Ojha et al., 2023) performs nearest-neighbor classification or linear probing on frozen CLIP features to study generalization across generative models. FatFormer (Liu et al., 2024) combines image- and frequency-domain forensic adaptation with image–text alignment supervision to learn forensic representations. Forensics Adapter (Cui et al., 2025) learns blending-boundary cues in a parallel adaptation branch that interacts with CLIP visual tokens. EFFORT (Yan et al., 2025) applies singular value decomposition to pretrained attention weights, freezes the dominant components, and adapts spectral residuals under orthogonality and singular-value constraints. Building on this representation foundation, SCORE further investigates semantic-conditioned primary prediction and controlled decision refinement.

3 PRELIMINARY

3.1 PROBLEM FORMULATION

A detector is trained on real/fake-labeled source data $\mathcal{D}_s$ and evaluated on unseen target domains $\mathcal{D}_t$, under shifts in generative mechanisms, content, capture conditions, and post-processing. The objective is reliable cross-domain prediction.

A central challenge in cross-domain forensics is the *context-dependent relevance of forensic evidence*. Texture, boundary, and frequency responses often jointly encode semantic content, imaging characteristics, and manipulation traces, while their discriminative value varies with pose, expression, scene structure, and image quality. A globally shared readout leaves context-dependent evidence weighting implicit in the encoder, which can limit adaptation when the target context distribution changes. The sufficiency of forensic evidence also varies across samples: dominant cues can resolve clear cases, while uncertain primary predictions motivate the controlled use of additional evidence.

These observations motivate SCORE (Sec. 4): organizing primary evidence by sample context and controlling subsequent refinement by primary prediction confidence.

3.2 SPECTRAL-RESIDUAL VISUAL ENCODING

We adopt the spectral-residual parameterization of EFFORT (Yan et al., 2025) as the shared visual encoding backbone. For a linear projection W in the self-attention layers of a pretrained Vision Transformer, the adapted weight is written as

$$W = W_{\text{main}} + \Delta W_\theta, \tag{1}$$

where W_{main} contains the frozen dominant singular components and ΔW_θ is a trainable low-rank spectral residual. This parameterization confines task-specific updates to a low-capacity residual subspace while preserving the dominant structure of the pretrained representation.

For an input image x_i, the spectrally adapted CLIP ViT-L/14 encoder (Radford et al., 2021) produces a shared visual representation

$$\mathbf{z}_i \leftarrow f_E(x_i), \qquad \mathbf{z}_i \in \mathbb{R}^D, \tag{2}$$

and a base classifier maps it to real/fake logits,

$$\mathbf{l}_i^{(0)} \leftarrow h_0(\mathbf{z}_i) = \left[l_{i,R}^{(0)}, l_{i,F}^{(0)}\right]^\top. \tag{3}$$

The spectral-residual encoder and base classifier form Baseline, which supplies shared features and base logits to both conditional modules. Its training objective and differences from EFFORT are detailed in Sec. 5.1.

4 SCORE

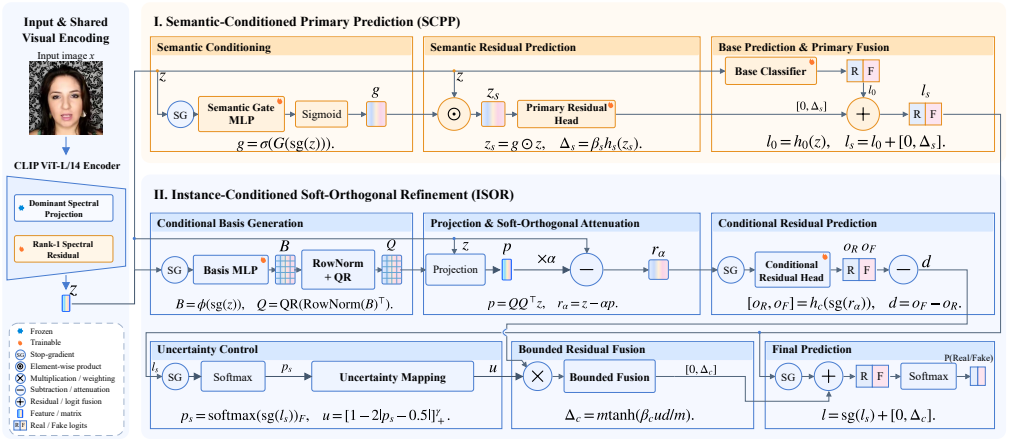


Figure 2: Overview of SCORE. SCPP forms an independently supervised, semantic-conditioned primary prediction; ISOR applies sample-conditioned soft-orthogonal refinement with an uncertainty-aware bounded correction. SG denotes stop-gradient.

4.1 Overview

SCORE applies two stages to the shared representation $\mathbf{z}_i$ (Fig. 2). SCPP forms an independently supervised primary prediction from semantic-conditioned evidence; ISOR uses a structured refinement representation and primary uncertainty to produce a bounded correction.

Let $\mathbf{e}_F = [0, 1]^\top$ denote the fake-logit selector, and let $\mathrm{sg}(\cdot)$ denote the stop-gradient operator. The two prediction stages are written as

$$\mathbf{l}_i^{(s)} \leftarrow \mathbf{l}_i^{(0)} + \mathbf{e}_F \Delta_{s,i}, \tag{4}$$

$$\mathbf{l}_i \leftarrow \mathrm{sg}(\mathbf{l}_i^{(s)}) + \mathbf{e}_F \Delta_{c,i}, \tag{5}$$

where $\Delta_{s,i}$ is the semantic-conditioned primary residual and $\Delta_{c,i}$ is the bounded refinement correction. The primary logits $\mathbf{l}_i^{(s)}$ are independently supervised; refinement updates their detached values. The final fake probability is

$$\pi_i \leftarrow \mathrm{softmax}(\mathbf{l}_i)_F. \tag{6}$$

4.2 Semantic-Conditioned Primary Prediction

The shared visual representation, carrying both semantic content and forensic information, supplies the conditioning signal. SCPP uses this signal to parameterize continuous channel modulation,

$$\mathbf{g}_i \leftarrow \sigma(G(\mathrm{sg}(\mathbf{z}_i))), \qquad \mathbf{z}_i^{(s)} \leftarrow \mathbf{g}_i \odot \mathbf{z}_i, \tag{7}$$

where G is a lightweight bottleneck network, σ denotes the sigmoid function, and $\mathbf{g}_i \in (0, 1)^D$. The detached $\mathbf{z}_i$ serves as the conditioning signal for gate generation, while the multiplicative feature branch retains gradients from the primary classification objective.

The modulated representation is mapped to a scalar fake-logit residual,

$$\delta_{s,i} \leftarrow h_s(\mathbf{z}_i^{(s)}), \qquad \Delta_{s,i} \leftarrow \beta_s(t)\delta_{s,i}, \tag{8}$$

where $\beta_s(t)$ controls the strength of the primary residual during training. The resulting primary prediction is

$$\mathbf{l}_i^{(s)} \leftarrow \mathbf{l}_i^{(0)} + \mathbf{e}_F \Delta_{s,i}. \tag{9}$$

Using the base logits as a reference, the scalar residual directly adjusts the real–fake logit margin according to the current semantic context. For a mini-batch of B samples, with $y_i = 0$ for real images and $y_i = 1$ for fake images, the primary prediction is trained with

$$\mathcal{L}_{\text{primary}} = \frac{1}{B}\sum_{i=1}^{B} \mathrm{CE}\big(\mathbf{l}_i^{(s)}, y_i\big). \tag{10}$$

This independent supervision makes $\mathbf{l}_i^{(s)}$ a complete prediction whose confidence can subsequently guide refinement.

4.3 Instance-Conditioned Soft-Orthogonal Refinement

Sample-conditioned soft-orthogonal representation. A conditional basis network ϕ generates k candidate directions from the detached shared representation,

$$\widetilde{\mathbf{B}}_i \leftarrow \mathrm{reshape}\left[\phi(\mathrm{sg}(\mathbf{z}_i))\right], \qquad \widetilde{\mathbf{B}}_i \in \mathbb{R}^{k\times D}. \tag{11}$$

After row-wise normalization, we obtain the orthonormal basis,

$$\mathbf{Q}_i \leftarrow \mathrm{QR}\left(\mathrm{RowNorm}(\widetilde{\mathbf{B}}_i)^\top\right), \qquad \mathbf{Q}_i^\top \mathbf{Q}_i = \mathbf{I}_k, \tag{12}$$

where $\mathbf{Q}_i \in \mathbb{R}^{D\times k}$ defines the low-dimensional conditional subspace for sample i. Here, $\mathrm{QR}(\cdot)$ returns the orthonormal $\mathbf{Q}$ factor of the reduced QR decomposition. The projection of the shared representation onto this subspace is

$$\mathbf{p}_i \leftarrow \mathbf{Q}_i \mathbf{Q}_i^\top \mathbf{z}_i. \tag{13}$$

To reduce basis-aligned responses while retaining part of the projected information, we apply soft attenuation,

$$\mathbf{r}_i^{(\alpha)} \leftarrow \mathbf{z}_i - \alpha \mathbf{p}_i, \qquad 0 < \alpha < 1. \tag{14}$$

Defining $\mathbf{r}_i^{\perp} \leftarrow \mathbf{z}_i - \mathbf{p}_i$, the same representation can be written as

$$\mathbf{r}_i^{(\alpha)} = \mathbf{r}_i^{\perp} + (1-\alpha)\mathbf{p}_i. \tag{15}$$

The orthogonal residual is fully preserved, while the basis-aligned component retains weight $1-\alpha$. This provides continuous control between the original representation and its orthogonal residual.

A conditional head operates on the detached refinement representation,

$$\mathbf{o}_i \leftarrow h_c\left(\mathrm{sg}(\mathbf{r}_i^{(\alpha)})\right) = \left[o_{i,R}, o_{i,F}\right]^\top, \qquad d_i \leftarrow o_{i,F} - o_{i,R}, \tag{16}$$

where d_i specifies the signed direction of the proposed fake-logit correction.

Uncertainty-aware bounded correction. We compute the fake probability of the detached primary prediction,

$$\pi_i^{(s)} \leftarrow \mathrm{softmax}\left(\mathrm{sg}(\mathbf{l}_i^{(s)})\right)_F, \tag{17}$$

and define the uncertainty weight as

$$u_i \leftarrow \left[1 - 2\left|\pi_i^{(s)} - \frac{1}{2}\right|\right]_+^{\gamma}. \tag{18}$$

Here, $[a]_+ = \max(a, 0)$ and $\gamma > 0$ controls the concentration of the uncertainty weight around $\pi_i^{(s)} = 0.5$. The weight reaches its maximum of 1 at $\pi_i^{(s)} = 0.5$ and decreases as the primary prediction becomes more confident.

The final refinement correction is

$$\Delta_{c,i} \leftarrow m \tanh\left(\frac{\beta_c(t) u_i d_i}{m}\right), \tag{19}$$

where $\beta_c(t)$ controls the training-stage strength of the refinement branch, u_i allocates the sample-wise refinement magnitude, d_i determines its direction, and $m > 0$ bounds the absolute correction. The correction is fused with the detached primary logits in Eq. (5). Appendix A.3 derives the corresponding bound on the primary-to-final probability change.

Training objective. The final prediction is supervised by

$$\mathcal{L}_{\text{final}} = \frac{1}{B}\sum_{i=1}^{B} \text{CE}\big(\mathbf{l}_i, y_i\big). \tag{20}$$

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{final}} + \mathcal{L}_{\text{primary}} + \mathcal{L}_{\text{SCPP}}^{\text{aux}} + \mathcal{L}_{\text{ref}}^{\text{aux}}, \tag{21}$$

where the auxiliary objectives regularize the semantic-conditioned primary prediction and the conditional refinement process. Auxiliary objectives and weights are given in Appendix A.4, gradient paths in Appendix A.2, and training schedules in Appendix A.5.

5 Experiments

5.1 Experimental Setup

Datasets and protocols. On UniversalFakeDetect (UFD) (Ojha et al., 2023), we train only on ProGAN (Karras et al., 2018) and evaluate 19 sources spanning GAN synthesis, face manipulation, low-level vision, perceptual synthesis, diffusion, and text-to-image generation.

Face models are trained on FaceForensics++ (FF++) c23 (Rössler et al., 2019). Cross-dataset evaluation covers CDF-v2 (Li et al., 2020b), DFD (Dufour & Gully, 2019), DFDC (Dolhansky et al., 2020), DFDCP (Dolhansky et al., 2019), and WDF (Zi et al., 2020); the DF40 (Yan et al., 2024b) cross-method protocol evaluates eight unseen manipulations. We additionally use DFo (Jiang et al., 2020) for qualitative and paired decision analyses under degradation. Full dataset names and test-source configurations are provided in Appendix B.1.

Baseline. Our Baseline retains the spectral-residual parameterization and base classification pathway of EFFORT (Yan et al., 2025): dominant singular components of the self-attention projections are frozen while low-rank spectral residuals are optimized. It uses binary classification supervision without the orthogonality regularizer $\mathcal{L}_{\text{orth}}$ and singular-value preservation regularizer $\mathcal{L}_{\text{ksv}}$. The resulting shared representation and base logits provide the common input to SCPP and ISOR. Baseline is used for internal comparisons, including component ablations and paired decision analyses; external comparison tables report published EFFORT results.

Implementation details. Our models use CLIP ViT-L/14 (Radford et al., 2021) with 224×224 inputs. SCPP and ISOR both use a bottleneck dimension of 128; ISOR uses $k = 8$, $\alpha = 0.25$, $\gamma = 1.25$, and $m = 0.25$. Both training settings use a learning rate of 2×10^{-4}. UFD uses AdamW for 18,000 optimizer steps with an effective batch size of 48; face models use Adam for 10 epochs with a batch size of 16 and are evaluated at the final epoch. Further preprocessing, sampling, augmentation, optimization, and scheduling details appear in Appendix B.1.

5.2 Comparison Experiments

On the local UFD evaluation subsets, SCORE achieves 99.70% mAP and 95.48% mAcc. Tables 1 and 2 provide published reference results under their reported settings. In the local component evaluation, SCORE improves over Baseline by 0.70 percentage points in mAP and 4.66 percentage points in mAcc (Table 5), with gains on DeepFakes, SITD, SAN, and Guided Diffusion (Table 4).

Cross-method face-forgery generalization. DF40 (Yan et al., 2024b) evaluates transfer to unseen manipulation methods under a content domain close to FF++. We further include EFFORT+TSRL (Lei et al., 2026), which reports 0.942 average AUC under the same protocol. Table 3 shows that SCORE reaches 0.972 average AUC and obtains the highest result on all eight manipulations.

5.3 Ablation Analysis

We evaluate four paired configurations under identical training and evaluation settings: Baseline, SCPP only (Baseline+SCPP), ISOR only (Baseline+ISOR), and the complete SCORE model (Base-

Table 1: AP (%) on UniversalFakeDetect. SCORE uses the local subsets specified in Appendix B.1; published reference results follow EFFORT (Yan et al., 2025) under their reported settings. † denotes results reported using official checkpoints or reproduced implementations. Bold and underline mark the highest and second-highest reported values, respectively.

Method	GAN						DeepFakes	Low-level		Perceptual loss		Guided	LDM			GLIDE			DALL-E	mAP
	ProGAN	CycleGAN	BigGAN	StyleGAN	GauGAN	StarGAN		SITD	SAN	CRN	IMLE		200	200 w/cfg	100	100/27	50/27	100/10		
CNN-Spot (Wang et al., 2020)	**100.00**	93.47	84.50	99.54	89.49	98.15	89.02	73.75	59.47	98.24	98.40	73.72	70.62	71.00	70.54	80.65	84.91	82.07	70.59	83.58
PatchForensics (Chai et al., 2020)	80.88	72.84	71.66	85.75	65.99	69.25	76.55	76.19	76.34	74.52	68.52	75.03	87.10	86.72	86.40	85.37	83.73	78.38	75.67	77.73
Co-occurrence (Nataraj et al., 2019)	99.74	80.95	50.61	98.63	53.11	67.99	59.14	68.98	60.42	73.06	87.21	70.20	91.21	89.02	92.39	89.32	88.35	82.79	80.96	78.11
Freq-Spec (Zhang et al., 2019)	55.39	**100.00**	75.08	55.11	66.08	**100.00**	45.18	47.46	57.12	53.61	50.98	57.72	77.72	77.25	76.47	68.58	64.58	61.92	67.77	66.21
F3Net† (Qian et al., 2020)	99.96	84.32	69.90	99.72	56.71	**100.00**	78.82	52.89	46.70	63.39	64.37	70.53	73.76	81.66	74.62	89.81	91.04	90.86	71.84	76.89
UniFD (Ojha et al., 2023)	**100.00**	98.13	94.46	86.66	99.25	99.53	91.67	78.54	67.54	83.12	91.06	79.24	95.81	79.77	95.93	93.93	95.12	94.59	88.45	90.14
LGrad† (Tan et al., 2023)	**100.00**	93.98	90.69	99.86	79.36	99.98	67.91	59.42	51.42	63.52	69.61	87.06	99.03	99.16	99.18	93.23	95.10	94.93	97.23	86.35
FreqNet† (Tan et al., 2024b)	99.92	99.63	96.05	99.89	99.71	98.63	**99.92**	94.42	74.59	80.10	75.70	96.27	96.06	**100.00**	62.34	99.80	99.78	96.39	77.78	91.95
NPR† (Tan et al., 2024a)	**100.00**	99.53	94.53	99.94	88.82	**100.00**	84.41	97.95	**99.99**	50.16	50.16	**98.26**	99.92	99.91	99.92	99.87	99.89	99.92	99.26	92.76
FatFormer† (Liu et al., 2024)	**100.00**	**100.00**	99.98	99.75	**100.00**	**100.00**	97.99	97.94	81.21	99.84	99.93	91.99	99.81	99.09	99.87	99.13	99.41	99.20	99.82	98.16
EFFORT (Yan et al., 2025)	**100.00**	**100.00**	**99.99**	99.77	**100.00**	**100.00**	98.95	97.53	97.53	**100.00**	**100.00**	95.39	**99.99**	99.89	**100.00**	99.87	99.92	99.98	99.96	99.41
SCORE	**100.00**	**100.00**	**99.99**	**99.95**	**100.00**	**100.00**	99.00	**99.22**	99.77	99.90	99.94	96.74	99.98	99.93	99.99	**99.94**	**99.93**	**100.00**	**99.97**	**99.70**

Table 2: Accuracy (%) on UniversalFakeDetect. SCORE uses class-balanced local subsets and a fixed threshold of 0.5, for which accuracy equals balanced accuracy. Published reference results follow EFFORT (Yan et al., 2025) under their reported sampling and threshold settings. † denotes results reported using official checkpoints or reproduced implementations. Bold and underline mark the highest and second-highest reported values, respectively.

Method	GAN						DeepFakes	Low-level		Perceptual loss		Guided	LDM			GLIDE			DALL-E	mAcc
	ProGAN	CycleGAN	BigGAN	StyleGAN	GauGAN	StarGAN		SITD	SAN	CRN	IMLE		200	200 w/cfg	100	100/27	50/27	100/10		
CNN-Spot (Wang et al., 2020)	99.99	85.20	70.20	85.70	78.95	91.70	53.47	66.67	48.69	86.31	86.26	60.07	54.03	54.96	54.14	60.78	63.80	65.66	55.58	69.58
PatchForensics (Chai et al., 2020)	75.03	68.97	68.47	79.16	64.23	63.94	75.54	75.14	75.28	72.33	55.30	67.41	76.50	76.10	75.77	74.81	73.28	68.52	67.91	71.24
Co-occurrence (Nataraj et al., 2019)	97.70	63.15	53.75	92.50	51.10	54.70	57.10	63.06	55.85	65.65	65.80	60.50	70.70	70.55	71.00	70.25	69.60	69.90	67.55	66.86
Freq-Spec (Zhang et al., 2019)	49.90	**99.90**	50.50	49.90	50.30	99.70	50.10	50.00	48.00	50.60	50.10	50.90	50.40	50.40	50.30	51.70	51.40	50.40	50.00	55.45
F3Net† (Qian et al., 2020)	99.38	76.38	65.33	92.56	58.10	**100.00**	63.48	54.17	47.26	51.47	51.47	69.20	68.15	75.35	68.80	81.65	83.25	83.05	66.30	71.33
UniFD (Ojha et al., 2023)	**100.00**	98.50	94.50	82.00	99.50	97.00	66.60	63.00	57.50	59.50	72.00	70.03	94.19	73.76	94.36	79.07	79.85	78.14	86.78	81.38
LGrad† (Tan et al., 2023)	99.84	85.39	82.88	94.83	72.45	99.62	58.00	62.50	50.00	50.74	50.78	77.50	94.20	95.85	94.80	87.40	90.70	89.55	88.35	80.28
FreqNet† (Tan et al., 2024b)	97.90	95.84	90.45	**97.55**	90.24	93.41	**97.40**	88.92	59.04	71.92	67.35	**86.70**	84.55	**99.58**	65.56	85.69	97.40	88.15	59.06	85.09
NPR† (Tan et al., 2024a)	99.84	95.00	87.55	96.23	86.57	99.75	76.89	66.94	**98.63**	50.00	50.00	84.55	97.65	98.00	98.20	96.25	97.15	97.35	87.15	87.56
FatFormer† (Liu et al., 2024)	99.89	99.32	99.50	97.15	99.41	99.75	93.23	81.11	68.04	69.45	69.45	76.00	98.60	94.90	98.65	94.35	94.65	94.20	**98.75**	90.86
EFFORT (Yan et al., 2025)	**100.00**	99.85	**99.60**	95.05	**99.60**	**100.00**	87.60	92.50	81.50	**98.90**	**98.90**	69.15	99.30	96.80	**99.45**	97.45	97.80	97.15	98.05	95.19
SCORE	**100.00**	99.45	98.90	95.35	99.30	**100.00**	90.65	**93.00**	93.00	89.05	89.05	77.15	**99.50**	96.60	99.20	**98.45**	**98.65**	**98.40**	98.50	**95.48**

Table 3: Video-level AUC on eight unseen manipulation methods in DF40. Results from F3Net through EFFORT follow EFFORT (Yan et al., 2025); TALL and EFFORT+TSRL follow TSRL (Lei et al., 2026). The SCORE average is computed before rounding individual results. Best and second-best results are shown in **bold** and underline, respectively.

Method	UniFace	BlendFace	MobileSwap	e4s	FaceDancer	FSGAN	InSwap	SimSwap	Avg.
F3Net (Qian et al., 2020)	0.809	0.808	0.867	0.494	0.717	0.845	0.757	0.674	0.746
SPSL (Liu et al., 2021)	0.747	0.748	0.885	0.514	0.666	0.812	0.643	0.665	0.710
SRM (Luo et al., 2021)	0.749	0.704	0.779	0.704	0.659	0.772	0.793	0.694	0.732
CORE (Ni et al., 2022)	0.871	0.843	0.959	0.679	0.774	0.958	0.855	0.724	0.833
RECCE (Cao et al., 2022)	0.898	0.832	0.925	0.683	0.848	0.949	0.848	0.768	0.844
SLADD (Chen et al., 2022)	0.878	0.882	0.954	0.765	0.825	0.943	0.879	0.794	0.865
SBI (Shiohara & Yamasaki, 2022)	0.724	0.891	0.952	0.750	0.594	0.803	0.712	0.701	0.766
TALL (Xu et al., 2023)	0.714	0.699	0.805	0.651	0.768	0.863	0.762	0.616	0.735
UCF (Yan et al., 2023)	0.831	0.827	0.950	0.731	0.862	0.937	0.809	0.647	0.824
IID (Huang et al., 2023)	0.839	0.789	0.888	0.766	0.844	0.927	0.789	0.644	0.811
LSDA (Yan et al., 2024a)	0.872	0.875	0.930	0.694	0.721	0.939	0.855	0.793	0.835
ProDet (Cheng et al., 2024)	0.908	0.929	0.975	0.771	0.747	0.928	0.837	0.844	0.867
CDFA (Lin et al., 2024)	0.762	0.756	0.823	0.631	0.803	0.942	0.772	0.757	0.781
EFFORT (Yan et al., 2025)	0.962	0.873	0.953	0.983	0.926	0.957	0.936	0.926	0.940
EFFORT + TSRL (Lei et al., 2026)	0.954	0.902	0.933	0.983	0.948	0.975	0.937	0.904	0.942
SCORE	**0.984**	**0.954**	**0.982**	**0.991**	**0.958**	**0.982**	**0.970**	**0.959**	**0.972**

line+SCPP+ISOR). Branch activation and loss counting for each configuration are specified in Appendix B.2.

Challenging cross-paradigm sources. On DeepFakes, SITD, SAN, and Guided Diffusion (Table 4), ISOR only improves AP and fixed-threshold balanced accuracy over Baseline. Adding SCPP further improves both metrics on all four sources and raises the 19-source mAP from 99.23% to 99.70% and mAcc from 93.89% to 95.48%.

Table 4: Component analysis on challenging UFD sources. Left: AP; right: fixed-threshold balanced accuracy (%). The final column averages all 19 UFD sources. Best and second-best results are shown in **bold** and underline, respectively.

(a) AP

Configuration	DeepFakes	SITD	SAN	Guided	mAP
Baseline	97.26	95.64	96.34	92.77	99.00
SCPP only	97.09	91.32	90.57	88.39	98.13
ISOR only	98.52	96.89	96.99	93.59	99.23
SCORE	**99.00**	**99.22**	**99.77**	**96.74**	**99.70**

(b) Balanced Accuracy

Configuration	DeepFakes	SITD	SAN	Guided	mAcc
Baseline	66.70	82.50	72.50	60.75	90.82
SCPP only	64.75	79.00	62.50	56.40	87.72
ISOR only	77.55	92.50	81.00	64.90	93.89
SCORE	**90.65**	**93.00**	**93.00**	**77.15**	**95.48**

Aggregate component effects. Table 5 summarizes all three protocols. Across the five cross-dataset targets, ISOR only and SCORE achieve average AUCs of 0.8986 and 0.8984, compared with 0.8927 for Baseline. Both reach 0.9723 on DF40. On UFD, the complete model achieves the highest mAP and mAcc, showing the clearest joint gain from the two stages.

Table 5: Aggregate component results across three generalization protocols. Cross-dataset AUC averages CDF-v2, DFD, DFDC, DFDCP, and WDF; DF40 AUC averages eight manipulation methods. Best and second-best results at the displayed precision, including ties, are shown in **bold** and underline, respectively.

Configuration	UFD mAP	UFD mAcc	Cross-dataset AUC	DF40 AUC
Baseline	99.00	90.82	0.8927	0.9661
SCPP only	98.13	87.72	0.8947	0.9698
ISOR only	99.23	93.89	**0.8986**	**0.9723**
SCORE	**99.70**	**95.48**	0.8984	**0.9723**

Cross-dataset error balance. Table 6 reports video-level EER across five target datasets. SCORE reduces the average from 0.1898 to 0.1873 and achieves the lowest mean EER among the four configurations, with reductions on DFD, DFDC, and WDF relative to Baseline.

Table 6: Video-level EER on five cross-dataset targets (lower is better). Avg. is the unweighted mean over the five listed datasets. Best and second-best results, including ties, are shown in **bold** and underline, respectively.

Configuration	CDF-v2	DFD	DFDC	DFDCP	WDF	Avg.
Baseline	**0.1404**	0.1791	0.2410	**0.1739**	0.2146	0.1898
SCPP only	0.1517	0.1736	**0.2359**	0.1870	0.2172	0.1931
ISOR only	**0.1404**	**0.1653**	0.2406	0.1957	**0.2020**	0.1888
SCORE	**0.1404**	**0.1653**	0.2393	0.1870	0.2045	**0.1873**

5.4 Qualitative and Paired Decision Analysis

5.4.1 Paired cross-domain score shifts

Figure 3 pairs Baseline and SCORE predictions for the same test videos from CDF-v2, DFDCP, WDF, and DFo. Across these four domains, 146 Baseline errors are corrected, including 138 fake false negatives; 99 new errors are introduced, including 94 real false positives. These paired differences between complete detectors include recovered forgeries and domain-dependent score shifts.

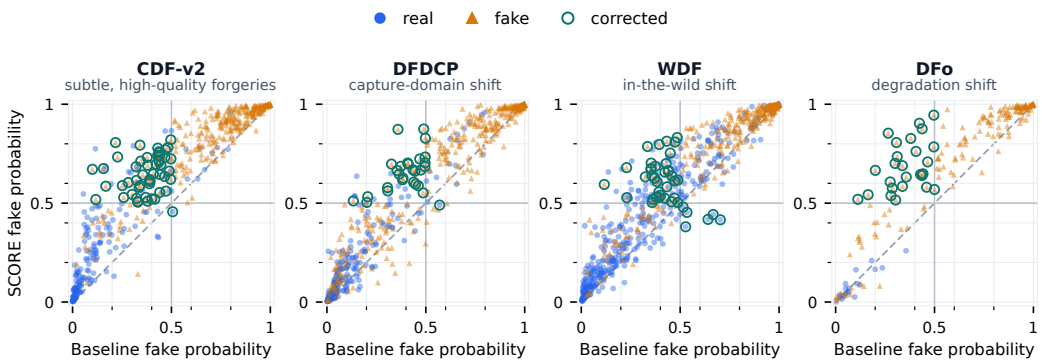


Figure 3: Paired video-level fake probabilities from Baseline and SCORE across four target domains. Each point represents one video; outlined points indicate corrected Baseline errors.

5.4.2 Cross-source threshold transfer

Figure 4 compares a fixed threshold of 0.5 with source-specific oracle thresholds on UFD. SCORE reduces the mean threshold-transfer gap across 19 sources from 6.25 to 3.10 percentage points, indicating more consistent score alignment under a shared decision threshold.

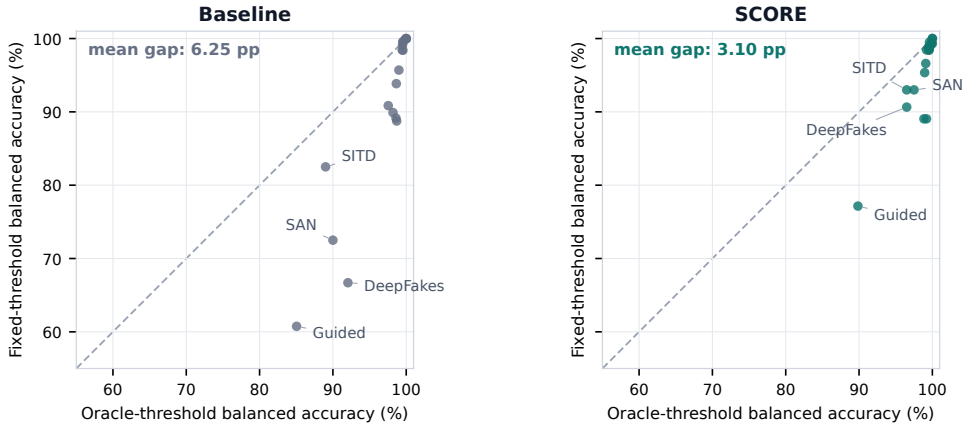


Figure 4: Cross-source threshold transfer on UniversalFakeDetect. Each point represents one test source, with oracle-threshold balanced accuracy on the horizontal axis and fixed-threshold (0.5) balanced accuracy on the vertical axis. The diagonal indicates zero threshold-transfer gap.

6 Conclusion

We presented SCORE, which combines semantic-conditioned primary prediction with uncertainty-aware bounded refinement for generalizable forgery detection. SCPP forms an independently supervised prediction, while ISOR uses a sample-conditioned soft-orthogonal view for refinement. Gradient isolation separates optimization roles, and the probability-change bound quantifies the final correction's influence within a fixed model. Across the evaluated UFD subsets and two face-forgery protocols, SCORE improves aggregate performance over the local baseline. The clearest joint benefit appears on UFD, while SCORE and ISOR-only achieve comparable AUC on the face protocols. Paired analyses reveal recovered false negatives and newly introduced errors, alongside a smaller UFD threshold-transfer gap. These results support context-dependent evidence formation with constrained decision revision.

AI USE STATEMENT

Generative AI tools assisted with drafting, translating, and polishing manuscript sections; retrieving and identifying related literature; and research ideation and execution, including methodological discussions and interpretation of experimental results. Additional uses included LaTeX preparation, diagram prototyping, consistency checks of method descriptions and experimental reporting, and assistance with deriving and presenting the probability-change bound in Appendix A.3. The authors retain full responsibility for the final manuscript, including its scientific claims, mathematical arguments, references, and reported results.

REFERENCES

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, 2022.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, 2020.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018.

Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*, 2022.

Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? In *NeurIPS*, 2024.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting CLIP for generalizable face forgery detection. In *CVPR*, 2025.

Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019.

Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Kh´ăc, Luke Melas, and Ritobrata Ghosh. DALL-E Mini, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Nick Dufour and Andrew Gully. Contributing data to deepfake detection research. Google Research Blog, September 2019. URL `https://research.google/blog/contributing-data-to-deepfake-detection-research/`.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, 2021.

Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *CVPR*, 2023.

Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. FreqDebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *CVPR*, 2025.

Zhanhe Lei, Zhongyuan Wang, Jikang Cheng, Baojin Huang, Yuhong Yang, Zhen Han, Chao Liang, and Dengpan Ye. Tutor-student reinforcement learning: A dynamic curriculum for robust deepfake detection. In *CVPR*, 2026.

Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse image synthesis from semantic layouts via conditional IMLE. In *ICCV*, 2019.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020a.

Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *CVPR Workshops*, 2019.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*, 2020b.

Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *ECCV*, 2024.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *CVPR*, 2021.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, 2024.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, 2021.

Lakshmanan Nataraj, Tajuddin Manhar Mohammed, B. S. Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H. Bappy, and Amit K. Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019.

Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. CORE: Consistent representation learning for face forgery detection. In *CVPR Workshops*, 2022.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *CVPR*, 2022.

Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. DiffusionFake: Enhancing generalization in deepfake detection via guided stable diffusion. In *NeurIPS*, 2024.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *CVPR*, 2023.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *AAAI*, 2024b.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. AltFreezing for more general video face forgery detection. In *CVPR*, 2023.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *ICCV*, 2023.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable deepfake detection. In *ICCV*, 2023.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *CVPR*, 2024a.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, and Li Yuan. DF40: Toward next-generation deepfake detection. In *NeurIPS*, 2024b.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *ICML*, 2025.

Wei Ye, Xinan He, and Feng Ding. Decoupling forgery semantics for generalizable deepfake detection. In *BMVC*, 2024.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *IEEE International Workshop on Information Forensics and Security*, 2019.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *ACM Multimedia*, 2020.

A METHOD SUPPLEMENT

This section provides the backbone parameterization, optimization roles, auxiliary objectives, and training schedules that complement the main method description.

A.1 SPECTRAL-RESIDUAL BACKBONE DETAILS

We use the vision encoder of CLIP ViT-L/14, which contains 24 Transformer blocks with hidden dimension $D = 1024$. Spectral-residual adaptation is applied to the query, key, value, and output projections of every self-attention block, covering 96 linear projections in total. For each projection, the pretrained weight is decomposed as

$$W = U_{1:D-1}\,\mathrm{diag}(S_{1:D-1})V_{1:D-1}^{\top} + U_r S_r V_r^{\top} = W_{\mathrm{main}} + \Delta W_\theta. \tag{22}$$

The first $D - 1 = 1023$ singular directions form the frozen dominant component W_{main}, while the remaining rank-1 component ΔW_θ is trainable. The original projection biases, LayerNorm layers, MLP blocks, patch/class/position embeddings, and post-layer normalization remain frozen.

The shared feature used by SCORE is the post-layer-normalized CLS representation before the CLIP visual projection,

$$\mathbf{z}_i \leftarrow \mathrm{LN}_{\mathrm{post}}\big(H_L^{(i)}[0]\big), \qquad \mathbf{z}_i \in \mathbb{R}^{1024}, \tag{23}$$

where $H_L^{(i)}[0]$ denotes the CLS token after the final Transformer block. The base classifier and SCORE-specific modules are optimized on top of this shared representation.

A.2 GRADIENT FLOW AND OPTIMIZATION ROLES

The stop-gradient operators in SCORE assign distinct optimization roles to the primary and refinement stages. The semantic gate is conditioned on $\mathrm{sg}(\mathbf{z}_i)$, while the gated feature $\mathbf{g}_i \odot \mathbf{z}_i$ retains gradients from the primary classification objective. The conditional refinement head operates on $\mathrm{sg}(\mathbf{r}_i^{(\alpha)})$, and the primary logits are detached before computing both the uncertainty weight and the final refinement.

Table 7 summarizes the resulting backward paths. The primary classification objective updates the shared encoder and SCPP, whereas the final classification objective directly optimizes the refinement head. The conditional-basis network receives gradients through the projection-energy range objective $\mathcal{L}_{\mathrm{range}}$.

Table 7: Effective gradient paths between loss terms and trainable parameter groups. A check mark indicates an active backward path.

Loss	Encoder	Base head	Gate	SCPP head	Basis MLP	Refine head
$\mathcal{L}_{\mathrm{final}}$						✓
$\mathcal{L}_{\mathrm{primary}}$	✓	✓	✓	✓		
$\mathcal{L}_{\mathrm{range}}$	✓				✓	
$\mathcal{L}_{\mathrm{logit}}, \mathcal{L}_{\Delta}, \mathcal{L}_{\mathrm{dir}}, \mathcal{L}_{\mathrm{pres}}$						✓
$\mathcal{L}_{\text{gate-sparse}}$			✓			
$\mathcal{L}_{\mathrm{coverage}}$	✓		✓			
$\mathcal{L}_{\mathrm{real}}, \mathcal{L}_{\text{sem-res}}$	✓		✓	✓		

A.3 PROBABILITY-CHANGE BOUND

For a fixed SCORE checkpoint, let $\xi_i^{(s)} = l_{i,F}^{(s)} - l_{i,R}^{(s)}$ denote the primary logit margin. The primary and final fake probabilities satisfy

$$\pi_i^{(s)} = \sigma(\xi_i^{(s)}), \qquad \pi_i = \sigma(\xi_i^{(s)} + \Delta_{c,i}). \tag{24}$$

For a scalar shift t, $\sup_a |\sigma(a + t) - \sigma(a)| = \tanh(|t|/4)$. Since $|\Delta_{c,i}| \le m$, it follows that

$$|\pi_i - \pi_i^{(s)}| \le \tanh\left(\frac{|\Delta_{c,i}|}{4}\right) \le \tanh\left(\frac{m}{4}\right). \tag{25}$$

At $m = 0.25$, the upper bound is approximately 0.0624. This bound concerns primary-to-final changes within the same checkpoint; comparisons between separately trained detectors also reflect differences in their learned representations and primary predictions.

A.4 Auxiliary Objectives

A.4.1 SCPP auxiliary objectives

For sample i and feature channel j, we compute the normalized feature energy as

$$\widehat{z}_{ij} \leftarrow \left(\frac{z_{ij}}{\|\mathbf{z}_i\|_2} \right)^2, \tag{26}$$

where $\mathcal{F}$ and $\mathcal{R}$ denote the fake and real samples in the current mini-batch, respectively. Using the unscaled primary residual $\delta_{s,i}$ defined in Sec. 4.2, the SCPP auxiliary terms are

$$\mathcal{L}_{\text{gate-sparse}} = \frac{1}{BD} \sum_{i,j} g_{ij}, \tag{27}$$

$$\mathcal{L}_{\text{coverage}} = \frac{1}{|\mathcal{F}|D} \sum_{i \in \mathcal{F}} \sum_{j} (1 - g_{ij}) \widehat{z}_{ij}, \tag{28}$$

$$\mathcal{L}_{\text{real}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \delta_{s,i}^2, \tag{29}$$

$$\mathcal{L}_{\text{sem-res}} = \frac{1}{B} \sum_{i} \delta_{s,i}^2. \tag{30}$$

The first two terms regulate channel modulation and the coverage of high-energy fake responses, while the latter two constrain the magnitude of the primary residual. Their weighted combination is

$$\begin{aligned} \mathcal{L}_{\text{SCPP}}^{\text{aux}} = a_s(t) \big(& 10^{-3} \mathcal{L}_{\text{gate-sparse}} + 10^{-3} \mathcal{L}_{\text{coverage}} \\ & + 10^{-3} \mathcal{L}_{\text{real}} + 10^{-4} \mathcal{L}_{\text{sem-res}} \big). \end{aligned} \tag{31}$$

A.4.2 Refinement auxiliary objectives

Using the projection $\mathbf{p}_i$ and orthogonal residual $\mathbf{r}_i^{\perp}$ defined in Sec. 4.3, the projection-energy ratio is computed as

$$\rho_i \leftarrow \frac{\|\mathbf{p}_i\|_2^2}{\|\text{sg}(\mathbf{z}_i)\|_2^2}, \tag{32}$$

and the projection-energy range objective is

$$\mathcal{L}_{\text{range}} = \frac{1}{B} \sum_{i} \left([0.04 - \rho_i]_+^2 + [\rho_i - 0.96]_+^2 \right). \tag{33}$$

This objective regulates the fraction of feature energy captured by the conditional projection. Orthogonality is guaranteed by construction: $\mathbf{Q}_i^\top \mathbf{Q}_i = \mathbf{I}_k$ implies $\mathbf{Q}_i^\top \mathbf{r}_i^{\perp} = \mathbf{0}$.

The conditional-logit and bounded-correction regularizers are

$$\mathcal{L}_{\text{logit}} = \frac{1}{2B} \sum_{i} \left(o_{i,R}^2 + o_{i,F}^2 \right), \tag{34}$$

$$\mathcal{L}_{\Delta} = \frac{1}{2B} \sum_{i} \Delta_{c,i}^2. \tag{35}$$

For directional supervision, we compute the detached primary uncertainty

$$w_i \leftarrow \left[1 - 2 \left| \pi_i^{(s)} - \frac{1}{2} \right| \right]_+, \tag{36}$$

and define the class-normalized directional objective as

$$\mathcal{L}_{\mathrm{dir}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{\sum_{i \in \mathcal{I}_c} w_i \, \ell_c(d_i)}{\max\big(\sum_{i \in \mathcal{I}_c} w_i, 1\big)}, \tag{37}$$

where $\mathcal{I}_R = \mathcal{R}$ and $\mathcal{I}_F = \mathcal{F}$ denote the real and fake sample index sets, respectively, and $\mathcal{C} = \{c \in \{R, F\} : |\mathcal{I}_c| > 0\}$ contains the classes present in the mini-batch. The class-specific losses are

$$\ell_F(d) = \mathrm{softplus}(-d), \qquad \ell_R(d) = \mathrm{softplus}(d). \tag{38}$$

To preserve confident primary predictions, the confidence weight is computed as

$$c_i \leftarrow 2 \left| \pi_i^{(s)} - \frac{1}{2} \right|, \tag{39}$$

and the preservation objective is

$$\mathcal{L}_{\mathrm{pres}} = \frac{\sum_i c_i D_{\mathrm{KL}}\Big(\mathrm{softmax}(\mathrm{sg}(\mathbf{l}_i^{(s)})) \parallel \mathrm{softmax}(\mathbf{l}_i)\Big)}{\max(\sum_i c_i, 1)}. \tag{40}$$

The effective refinement auxiliary objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{ref}}^{\mathrm{aux}} = a_c(t)\big(&0.001\mathcal{L}_{\mathrm{range}} + 0.0002\mathcal{L}_{\mathrm{logit}} \\ &+ 0.0015\mathcal{L}_{\Delta} + 0.001\mathcal{L}_{\mathrm{dir}} + 0.02\mathcal{L}_{\mathrm{pres}}\big). \end{aligned} \tag{41}$$

The implementation also evaluates $(Bk)^{-1} \sum_i \|\mathbf{Q}_i^\top \mathbf{r}_i^\perp\|_2^2$ with weight $0.005 a_c(t)$; this quantity vanishes in exact arithmetic and is omitted from the analytical objective.

A.5 Training Schedules and Hyperparameters

SCPP and the refinement stage are progressively introduced during training using the reference-step ramp

$$\eta(t; w, r) = \begin{cases} 0, & t < w, \\ \min\left(1, \dfrac{t - w + 1}{r}\right), & t \geq w, \end{cases} \qquad t \leftarrow t_{\mathrm{opt}} \frac{B_{\mathrm{train}}}{B_{\mathrm{ref}}}, \tag{42}$$

where t_{opt} denotes the optimizer-step index.

The primary and refinement coefficients are defined as

$$\beta_s(t) = 0.15\, \eta(t; 540, 1080), \qquad a_s(t) = \eta(t; 540, 1080), \tag{43}$$

$$\beta_c(t) = 0.035\, \eta(t; 1080, 1080), \qquad a_c(t) = \eta(t; 1080, 1080). \tag{44}$$

For face experiments, $B_{\mathrm{train}} = B_{\mathrm{ref}} = 16$, so the reference-step index equals the optimizer-step index. For UFD, the effective batch size is 48, and the reference-step index advances by three units per optimizer step. At inference time, the residual scales are fixed at $\beta_s = 0.15$ and $\beta_c = 0.035$.

Table 8 summarizes the remaining fixed hyperparameters.

Table 8: SCORE hyperparameters used in all reported experiments.

Hyperparameter	Value
Feature dimension D	1024
SCPP bottleneck dimension	128
Basis-network bottleneck dimension	128
Conditional subspace dimension k	8
Soft attenuation coefficient α	0.25
Uncertainty exponent γ	1.25
Maximum SCPP residual scale	0.15
Conditional refinement scale	0.035
Maximum correction magnitude m	0.25
Projection-energy range	[0.04, 0.96]
Residual-head weights and biases	Zero

B EXPERIMENTAL SUPPLEMENT

B.1 ADDITIONAL EXPERIMENTAL DETAILS

UFD test sources. The 19 UniversalFakeDetect test sources (Ojha et al., 2023) are ProGAN (Karras et al., 2018), CycleGAN (Zhu et al., 2017), BigGAN (Brock et al., 2019), StyleGAN (Karras et al., 2019), GauGAN (Park et al., 2019), StarGAN (Choi et al., 2018), DeepFakes (Rössler et al., 2019), SITD (Chen et al., 2018), SAN (Dai et al., 2019), CRN (Chen & Koltun, 2017), IMLE (Li et al., 2019), Guided Diffusion (Dhariwal & Nichol, 2021), three LDM settings (Rombach et al., 2022), three GLIDE settings (Nichol et al., 2022), and DALL-E Mini (Dayma et al., 2021), labeled DALL-E in the result tables.

UFD training and preprocessing. The local ProGAN training set covers 20 semantic categories and contains 720,119 PNG images: 360,059 real and 360,060 generated images. Small images are expanded by tiling before cropping, using a size threshold of 256 pixels. Training applies random 224×224 cropping, random horizontal flipping, and CLIP normalization; evaluation uses center 224×224 cropping and the same normalization. Interpolation-based resizing is disabled, and UFD training uses no additional JPEG-compression or blur augmentation. Evaluation uses the checkpoint at optimizer step 18,000.

UFD evaluation subsets. The local evaluator requests 1,000 real and 1,000 generated images per source. If either class contains fewer than 1,000 images, the sample count is reduced to 100 per class. Accordingly, 17 sources are evaluated using 2,000 images each, while SITD and SAN use 200 images each. Sampling uses seed 0. These settings apply to the local component evaluations; published reference results retain the evaluation settings of their sources.

Face datasets and manipulation methods. The five cross-dataset targets are Celeb-DF-v2 (CDF-v2) (Li et al., 2020b), DeepFakeDetection (DFD) (Dufour & Gully, 2019), Deepfake Detection Challenge (DFDC) (Dolhansky et al., 2020), DFDC Preview (DFDCP) (Dolhansky et al., 2019), and WildDeepfake (WDF) (Zi et al., 2020). Additional qualitative and paired decision analyses use DeeperForensics-1.0 (DFo) (Jiang et al., 2020). The DF40 (Yan et al., 2024b) cross-method evaluation includes UniFace, BlendFace, MobileSwap, e4s, FaceDancer, FSGAN, InSwap, and SimSwap.

Metrics and comparisons. For each local UFD source, we report AP and balanced accuracy, with the latter evaluated at threshold 0.5. Their unweighted averages over the 19 sources are denoted by mAP and mAcc, respectively. Each local source subset contains equal numbers of real and generated images, so balanced accuracy equals overall accuracy. Face-forgery evaluation reports video-level ROC-AUC and EER. Published UFD and DF40 reference results are taken from EFFORT (Yan et al., 2025), with additional DF40 results from TSRL (Lei et al., 2026). Published values, including reported averages, are retained as given by their sources.

Face preprocessing and frame selection. Face experiments use preprocessed face crops. Each crop is bicubically resized to 224×224 and normalized with the CLIP mean $(0.48145466, 0.4578275, 0.40821073)$ and standard deviation $(0.26862954, 0.26130258, 0.27577711)$. The data loader follows the frame order stored in the metadata and selects the first N frame entries for each video, with $N = 8$ for training and cross-dataset evaluation and $N = 32$ for DF40 evaluation. Qualitative and paired decision analyses on DFo use the processed subset of 20 real and 185 fake videos with valid frame entries.

Optimization details. For face experiments, Adam uses $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and weight decay 5×10^{-4}. All configurations use a fixed seed of 1024, and neither AMP nor gradient clipping is applied. For UFD, AdamW uses $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and zero weight decay. A micro-batch size of 16 with gradient accumulation over three steps gives the effective batch size of 48 reported in the main paper. UFD uses seed 0. The experiments are conducted on a server equipped with four NVIDIA GeForce RTX 3090 24GB GPUs, with each training run using a single GPU.

Face-training augmentation. Face training applies horizontal flipping ($p = 0.5$), brightness/contrast adjustment ($p = 0.5$, magnitude 0.2), hue/saturation/value augmentation ($p = 0.3$), and

JPEG compression with quality 40–100 ($p = 0.1$). Gaussian noise, motion blur, CLAHE, channel shuffle, and cutout are each applied with probability 0.1, while random gamma and glass blur are each applied with probability 0.3. Face-forgery evaluation uses only resizing, tensor conversion, and CLIP normalization.

Video aggregation and threshold diagnostics. For face-forgery evaluation, frame-level fake probabilities are averaged within each video,

$$\bar{\pi}_v = \frac{1}{|\mathcal{I}_v|} \sum_{i \in \mathcal{I}_v} \mathrm{softmax}(\mathbf{l}_i)_F, \tag{45}$$

where $\mathcal{I}_v$ denotes the selected frame set of video v. Video-level ROC-AUC and EER are computed from $\bar{\pi}_v$.

For each local UFD source subset s, let $\tau_s^\star$ denote the threshold that maximizes accuracy on that subset. The threshold-transfer gap used in Fig. 4 is

$$G_s = \mathrm{BAcc}_s(\tau_s^\star) - \mathrm{BAcc}_s(0.5), \qquad \overline{G} = \frac{1}{19} \sum_s G_s. \tag{46}$$

The source-specific oracle threshold is used only for this diagnostic analysis.

B.2 Complete Component Results

Ablation configurations. Baseline uses the base prediction with one classification loss. For SCPP only, the primary prediction is also the final output; its classification loss is counted once, SCPP auxiliary objectives are active, and ISOR is disabled. For ISOR only, $\mathbf{l}_i^{(s)} \leftarrow \mathbf{l}_i^{(0)}$, and uncertainty is computed from the detached base prediction; primary and final classification losses and refinement auxiliaries are active, while SCPP is disabled. The complete SCORE model uses both classification losses and both groups of auxiliary objectives. Disabled branches contribute neither auxiliary losses nor scheduled residuals; their parameters are frozen in the face-forgery implementation and omitted from the UFD model.

B.2.1 Cross-dataset AUC

Table 9 reports the per-target AUC across five datasets. ISOR only achieves the highest average AUC of 0.8986; SCORE reaches 0.8984, compared with 0.8927 for Baseline.

Table 9: Complete component results on five cross-dataset targets in terms of video-level AUC. Avg. is the unweighted mean over the five listed datasets. Best and second-best results, including ties, are shown in **bold** and underline, respectively.

Configuration	CDF-v2	DFD	DFDC	DFDCP	WDF	Avg.
Baseline	0.9422	0.9162	0.8428	0.8951	0.8673	0.8927
SCPP only	0.9341	0.9173	**0.8441**	**0.9067**	0.8714	0.8947
ISOR only	**0.9446**	**0.9236**	0.8439	0.9018	**0.8792**	**0.8986**
SCORE	<u>0.9438</u>	<u>0.9235</u>	<u>0.8440</u>	<u>0.9019</u>	<u>0.8790</u>	<u>0.8984</u>

B.2.2 DF40 component results

Table 10 reports AUC for the eight unseen manipulations. ISOR only and SCORE both achieve an average AUC of 0.9723 at the displayed precision, while SCPP only provides its largest gains over Baseline on BlendFace and SimSwap.

Table 10: Complete component AUC results on eight unseen manipulations in DF40. Best and second-best results are shown in **bold** and underline, respectively.

Configuration	UniFace	BlendFace	MobileSwap	e4s	FaceDancer	FSGAN	InSwap	SimSwap	Avg.
Baseline	0.9799	0.9371	0.9746	**0.9914**	0.9506	**0.9866**	0.9669	0.9418	0.9661
SCPP only	0.9823	**0.9569**	0.9789	0.9905	0.9421	0.9839	**0.9697**	0.9544	0.9698
ISOR only	0.9835	0.9544	**0.9821**	0.9909	**0.9579**	0.9818	0.9696	0.9582	**0.9723**
SCORE	**0.9837**	0.9539	**0.9821**	0.9909	0.9577	0.9816	0.9696	**0.9586**	**0.9723**

B.3 Additional Qualitative Results

Figure 5 illustrates sample-level outcomes of SCORE relative to Baseline: recovered false negatives, remaining false negatives, and selected newly introduced false positives. These examples complement the paired score analysis of the complete detectors.

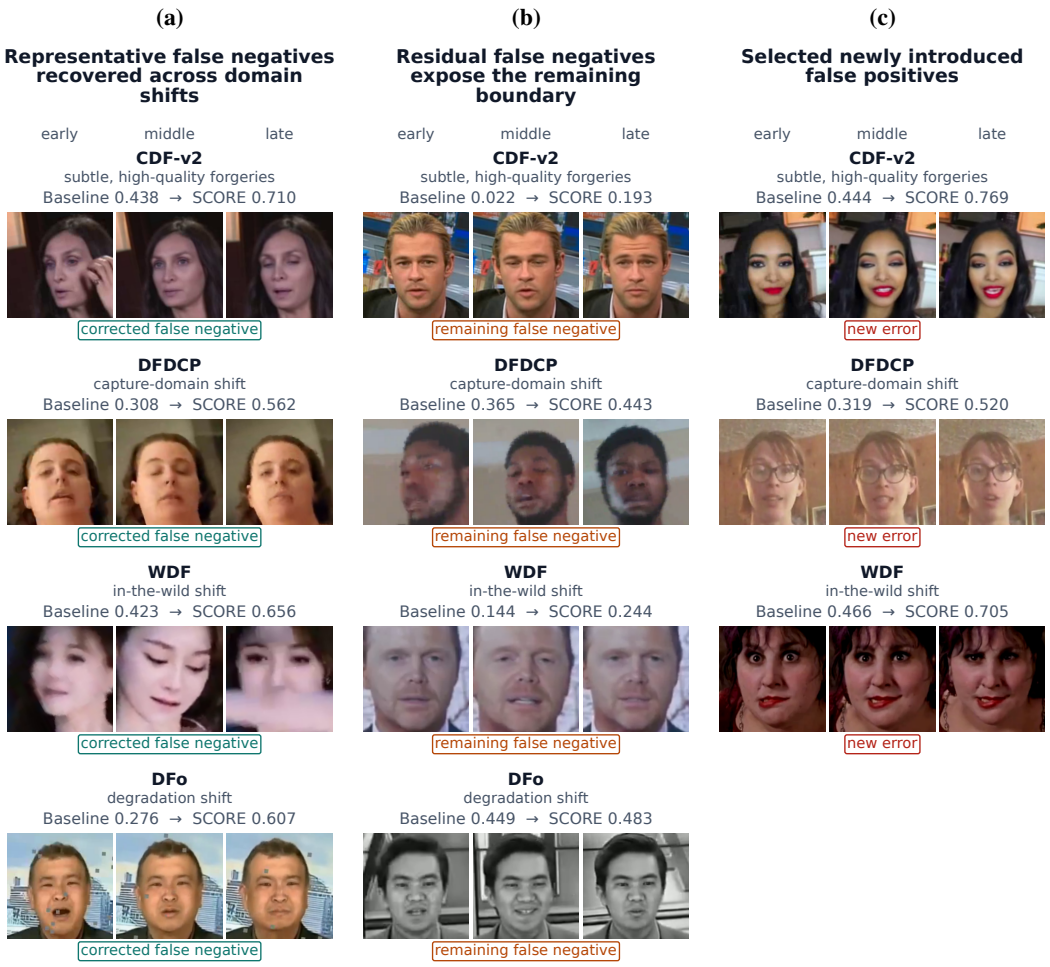


Figure 5: Additional qualitative results: (a) recovered false negatives, (b) remaining false negatives, and (c) selected newly introduced false positives. Each row shows three frames from one video and paired fake probabilities (Baseline → SCORE). Samples are selected deterministically.