

Face-Aligned Image Ensembles and LFCC Audio Transformers for ImageCLEF 2026 Deepfake Detection

Qingbiao Hu¹, Zhongyuan Han*, Ziran Su, Jinghong Peng, Weibin Huang and Weilong Huang

¹Foshan University, Foshan, China

Abstract

The ImageCLEF 2026 Deepfake Detection and Generation Task evaluates binary detection of real and deepfake image and audio content. For the Image and Audio Detection subtasks, we submitted a two-branch system. The image branch applies face alignment and averages the scores of CLIP, GenD, and NPR detectors. The audio branch extracts LFCC representations and averages ViT-Tiny and ResNet18 outputs. On DF40-derived validation data, the image ensemble obtained 0.855 accuracy compared with 0.825 for the CLIP baseline, and face alignment increased accuracy from 0.562 to 0.638 in a padded-image simulation. The audio ViT-Tiny model obtained 96.65% AUC on a local split built from LJSpeech and WaveFake. In the official evaluation, the final scores were 0.6038 for image detection and 0.5595 for audio detection. The official image subset scores indicate false positives on unseen real images, while the audio result shows limited transfer from the local single-speaker and vocoder data.

Keywords

deepfake detection, ImageCLEF 2026, face alignment, audio detection, CLIP, LFCC

1. Introduction

Deepfakes include manipulated or generated faces and synthetic speech produced by face-swapping, reenactment, image synthesis, text-to-speech, and voice-conversion systems. Detection methods formulate the problem as binary classification and learn visual or acoustic traces that separate real from generated content. For images, these traces can include semantic inconsistencies and local artifacts introduced by synthesis or upsampling. For audio, they can appear as spectral and temporal patterns associated with acoustic models, vocoders, recording channels, or compression.

A central difficulty is generalization to data that differ from the training distribution. Image detectors can learn manipulation-specific artifacts, source-camera characteristics, compression patterns, or preprocessing cues instead of features that transfer to unseen generators. Cross-dataset evaluations in DeepfakeBench and DF40 show that performance can vary with the manipulation family and dataset source [1, 2]. Audio countermeasures face related changes in speakers, languages, channels, codecs, and synthesis systems; ASVspoof 2021 therefore evaluated speech deepfake detection without matched training and development data [3].

ImageCLEF 2026 hosts multimodal retrieval and analysis tasks across several domains [4]. We participated in the Image and Audio Detection subtasks of the Deepfake Detection and Generation Task. Both subtasks require a hard binary label for every test item, with 0 denoting real content and 1 denoting deepfake content. No labeled task-specific training set was supplied, and the blind evaluation includes organizer data and deepfakes contributed by participants in the generation subtask [5, 6]. This setting combines manipulation shift with differences between external training data and the official test data.

Our image system aligns faces and averages the outputs of a DF40-finetuned CLIP classifier, GenD, and NPR. The audio system converts each waveform to a fixed LFCC representation and averages ViT-Tiny and ResNet18 probabilities. The official final scores were 0.6038 for image detection and 0.5595 for audio detection. The subset results indicate that real-image false positives remained a limitation,

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

while the gap between local and official audio results indicates limited transfer from LJSpeech and WaveFake to the blind set.

The remainder of the paper is organized as follows. Section 2 reviews image and audio detection methods and the generalization problem. Section 3 defines the task and describes both submitted systems. Section 4 reports the data, implementation, development experiments, official results, and their interpretation. Section 5 concludes the paper.

2. Related Work

2.1. Image Deepfake Detection

Early face-forensics benchmarks commonly train convolutional classifiers on spatial image content. FaceForensics++ contains manipulated videos produced by several face-swapping and reenactment methods [7], while Celeb-DF was designed to reduce obvious synthesis artifacts and provide a more challenging cross-dataset test [8]. DeepfakeBench standardizes datasets, detectors, and evaluation protocols [1]. DF40 extends this direction by organizing 40 face-forgery techniques across face swapping, reenactment, entire-face synthesis, and editing [2].

Pretrained vision-language models provide another representation for fake-image detection. CLIP learns transferable visual features through image-text pretraining [9]. UnivFD uses a frozen CLIP vision encoder and a linear classifier, seeking features that transfer across generative models rather than fitting one generator’s artifacts [10]. Our DF40 CLIP baseline instead uses a two-class head adapted to face deepfake data.

Artifact-based and parameter-efficient methods provide different cues. NPR represents neighboring-pixel relationships to expose structural traces associated with upsampling in generative networks [11]. GenD retains the CLIP vision backbone, tunes its Layer Normalization parameters, normalizes the classification feature, and applies metric-learning objectives and latent-space augmentation during training [12]. We evaluate CLIP, GenD, NPR, and UnivFD because they differ in training data and in the visual evidence used for classification.

2.2. Audio Deepfake Detection

Audio countermeasures use either engineered acoustic features or representations learned directly from waveforms. Cepstral features such as LFCC summarize short-time spectral structure and can be classified with statistical models or neural networks. The ASVspoof series provides common protocols and reports EER and detection-cost metrics; its 2021 edition added a speech deepfake task with channel and compression variability [3]. WaveFake contains nine generated-speech sets from five model architectures and two languages, together with signal-processing baselines [13].

Neural audio countermeasures model local and long-range structure in different ways. Residual convolutional networks aggregate local time-frequency patterns [14], while transformers divide a representation into patches and model interactions through self-attention [15]. Dedicated systems such as AASIST combine spectral and temporal graph-attention branches over a learned waveform representation [16]. Our audio submission uses normalized LFCC inputs with ViT-Tiny and ResNet18 rather than a dedicated end-to-end anti-spoofing architecture.

2.3. Generalization Across Domains

Generalization depends on both the detector and the construction of the evaluation data. For face images, changes in the real-image source, manipulation method, resolution, compression, and face scale can change the score distribution. For audio, speaker identity, language, recording channel, codec, duration, and synthesis architecture can play the same role. Broad benchmarks and paired real-fake training reduce some source-specific shortcuts, but no single training collection covers all future generators [2, 12, 3]. Our systems address two limited aspects of this problem: face alignment

reduces an image-scale mismatch, and score averaging combines models with different error patterns. The official results are used to identify the distribution shifts that remain.

3. Method

3.1. Task Formulation

Let x_i^m be a test item from modality $m \in \{I, A\}$, where I denotes an image and A denotes an audio waveform. The goal is to predict $y_i^m \in \{0, 1\}$, where 0 indicates real content and 1 indicates deepfake content. For each modality, our system estimates a fake score $p_m(x_i^m)$ and produces a required hard label

$$\hat{y}_i^m = \mathbb{I}[p_m(x_i^m) > \tau_m]. \quad (1)$$

The image branch uses $\tau_I = 0.5$; the audio implementation takes the argmax of its two-class fused probability, which is equivalent to a 0.5 decision for the fake class.

3.2. Detection System

Figure 1 summarizes the system. The thick-bordered blocks denote preprocessing and score-fusion steps.

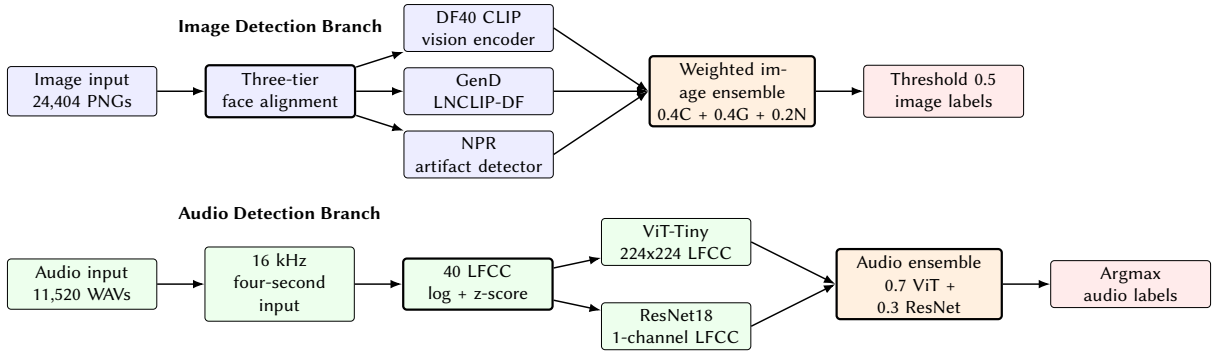


Figure 1: Architecture of the submitted detection system. Thick-bordered blocks denote preprocessing and score-fusion steps. Both branches produce hard labels: 0 for real and 1 for deepfake.

3.2.1. Image Detection System

Face alignment. The final pipeline starts by converting each input image into an aligned 256 by 256 face crop. We used a three-tier cropper:

1. dlib HOG face detection followed by an 81-point landmark predictor and the same 5-point similarity transform used in the DF40 preprocessing pipeline, with scale 1.3 and output size 256 by 256;
2. OpenCV Haar cascade detection as a fallback when dlib failed, followed by a square crop around the detected face and bicubic resize;
3. full-image bicubic resize as a last resort when neither face detector found a face.

The cropper is applied only to CLEF-style full-frame inputs. For already aligned DF40 validation images, applying the cropper a second time was harmful, as shown in Section 4.2.3.

After cropping, each detector receives its own expected resize and normalization. The CLIP baseline and GenD receive 224 by 224 inputs; NPR receives 256 by 256 inputs. We used the CLIP native mean and standard deviation for GenD and UnivFD-style models, the DF40 baseline normalization with mean and standard deviation equal to 0.5 for the CLIP baseline, and ImageNet normalization for NPR.

Base detectors. We evaluated CLIP, GenD, NPR, and UnivFD during development. The final submission used the first three models; UnivFD was not included because it transferred poorly to face-swap and reenactment methods.

The CLIP baseline uses the large CLIP vision encoder with 14-pixel patches and a two-class linear head; subsequent references call it the CLIP encoder. The transformer divides a 224 by 224 image into a 16 by 16 grid of tokens, each covering a 14 by 14 pixel patch, and uses its global representation for binary classification. We used the checkpoint fine-tuned in the DF40 framework. This model supplies a semantic representation, but its output can also reflect the source distribution used during fine-tuning.

GenD uses the same CLIP encoder but changes how it is adapted. LNCLIP-DF tunes the Layer Normalization parameters while leaving most of the encoder fixed, applies L2 normalization to the classification feature, and uses alignment, uniformity, and latent-space augmentation during training [12]. We used the released TorchScript checkpoint and its two-class output.

NPR is an artifact-oriented detector. It transforms the RGB image into neighboring-pixel-relation residuals and classifies them with a truncated ResNet-50, so its decision is based on local structural patterns associated with generative upsampling rather than CLIP semantics [11]. UnivFD, used only as a comparison, applies a linear fake-image classifier to frozen CLIP features [10].

Fusion and decision rule. Each base detector outputs a probability of the fake class. For the submitted system, we used the weighted average

$$p_{\text{ens}} = 0.4p_{\text{CLIP}} + 0.4p_{\text{GenD}} + 0.2p_{\text{NPR}}. \quad (2)$$

The submitted hard label is 1 when $p_{\text{ens}} > 0.5$ and 0 otherwise. The weights were selected from the validation comparisons in Section 4.2.1; NPR received lower weight because its results varied across source groups.

3.2.2. Audio Detection System

LFCC preprocessing. For audio detection, each waveform was loaded with `librosa`, resampled to 16 kHz, and either truncated or zero-padded to four seconds. We then extracted 40 LFCC coefficients using `torchaudio`, with `n_fft=512`, `hop_length=160`, and centered windows. The LFCC tensor was transformed with $\log_{10}(x + 10^{-9})$, normalized per sample with z-score normalization, and sanitized by clipping non-finite values. LFCC uses a linearly spaced filter bank and a cepstral transform to summarize the short-time spectrum. The resulting single-channel time-frequency tensor has a fixed shape for all utterances and is used as the input to both classifiers.

ViT-Tiny classifier. The primary audio model was a ViT-Tiny classifier based on `vit_tiny_patch16_224` [15]. Single-channel LFCC inputs were bilinearly resized to 224 by 224 and repeated across three channels to match the ImageNet-pretrained ViT input format. The model splits this representation into 16 by 16 patches, applies self-attention across the time-frequency plane, and feeds the global token to a two-class head. Starting from an existing ViT checkpoint, training used five epochs of head-only warmup with Adam at learning rate 10^{-3} , followed by five epochs of full fine-tuning with AdamW at learning rate 10^{-7} , weight decay 0.01, batch size 32, and gradient clipping at norm 1.0.

ResNet18 classifier and fusion. We also trained a ResNet18 classifier [14] on the same LFCC representation. Its residual convolutional blocks aggregate local patterns in the LFCC plane. The first convolution was modified to accept one channel, the final layer was replaced by a two-class classifier, and the network was trained from scratch for five epochs with AdamW, learning rate 10^{-4} , weight decay 0.01, batch size 32, and gradient clipping at norm 1.0.

The submitted audio prediction used weighted probability averaging:

$$p_{\text{audio}} = 0.7p_{\text{ViT}} + 0.3p_{\text{ResNet}}. \quad (3)$$

The final label was the argmax of the averaged two-class probability vector. The ViT receives the larger weight because it was the primary locally evaluated model; ResNet18 supplies a second score based on convolutional features.

4. Experiments

4.1. Experimental Setup

4.1.1. Experimental data

The official image blind set contains 24,404 PNG files. All released images are 256 by 256 pixels, but their face scale varies: some images are face-centered, while others contain a smaller face within the frame. The labels were hidden during development. The organizers later reported scores for baseline deepfakes, public organizer real data, a secret verified real subset, and participant-generated deepfakes.

We used DF40 [2] with DeepfakeBench-style evaluation code [1] for image model selection. DF40 covers face swapping, reenactment, entire-face synthesis, and editing methods. Our main validation benchmark contains 250,802 unique frames from 80 DF40-derived datasets, grouped in the analysis as FaceForensics++-domain (ff), Celeb-DF-domain (cdf), and standalone sources. For fusion evaluation, we retained method-level records, yielding 337,407 entries because the same real frame can serve as the control for more than one manipulation method. The CLIP baseline uses DF40 resources based on the FaceForensics++ domain; Celeb-DF-derived samples were used for out-of-domain evaluation rather than for that baseline’s training. We checked the local manifests for filename overlap between the training data and the CLEF blind set and found none.

The official audio blind set contains 11,520 variable-duration WAV files. For local training and evaluation, LJSpeech-1.1 supplies 13,100 English clips from one speaker, totaling approximately 24 hours [17]. WaveFake supplies generated speech from multiple vocoder and synthesis configurations; the original collection contains nine sample sets from five model architectures and two languages [13]. Our combined metadata contains 147,366 labeled examples. The training split contains 117,892 examples (10,499 real and 107,393 fake), the validation split contains 14,737 (1,359 real and 13,378 fake), and the local test split contains 14,737 (1,242 real and 13,495 fake). Approximately 91% of each local split is fake, so local accuracy and threshold selection are affected by class imbalance.

4.1.2. Data processing

For image blind-test inference, each frame is processed by the three-tier face alignment pipeline in Section 3.2.1 before model-specific resizing and normalization. DF40 validation frames already match the aligned-face training format and are therefore evaluated without recropping, except in the explicit preprocessing ablations. On the 24,404 official images, dlib alignment was used for 20,309 images (83.2%), the Haar fallback for 2,558 (10.5%), and full-image resizing for 1,537 (6.3%). CLIP and GenD receive 224 by 224 tensors, whereas NPR receives 256 by 256 tensors; each model uses the normalization associated with its checkpoint. For audio, waveforms are decoded as mono, resampled to 16 kHz, truncated or zero-padded to 64,000 samples, and converted to normalized 40-dimensional LFCC tensors as described in Section 3.2.2.

4.1.3. Evaluation metrics

The challenge submissions require one binary prediction for each filename, with `prediction=0` for real and `prediction=1` for deepfake. The task documentation lists accuracy, precision, recall, and F1 as detection metrics [5, 6]. The official results supplied to participants report accuracy on each evaluation subset. The primary final score is

$$S = 0.1A_{\text{base}} + 0.1A_{\text{real}} + 0.4A_{\text{realGT}} + 0.4A_{\text{part,w}}, \quad (4)$$

where the last term is participant-deepfake accuracy weighted by the realism score from the generation task. The unweighted participant-deepfake accuracy is reported separately but is not an additional term in S . For image development we additionally report AUC, real-class accuracy, fake-class accuracy, and balanced accuracy to identify failures hidden by uneven source distributions. For local audio evaluation, we report AUC, equal error rate (EER), and minimum tandem detection cost function (min-tDCF), alongside the official blind-set score returned by the evaluation server.

4.1.4. Baseline methods

The primary image baseline is the DF40-finetuned CLIP classifier without fusion. We compare it with the individual GenD and NPR models, an evaluated but excluded UnivFD candidate, unweighted and weighted probability averaging, voting, maximum and minimum fake-score rules, and a two-model CLIP and GenD fusion. For audio, ViT-Tiny over LFCC serves as the single-model baseline and the submitted method adds ResNet18 probability fusion. The image models were evaluated in PyTorch 2.4.1 with their published or DF40 checkpoints. The submission pipeline used batches of 48 images and applied the model-specific resizing and normalization described in Section 3.2.1.

4.2. Experimental Results

4.2.1. Single-model image comparison

We group DF40-derived validation data into FaceForensics++-domain data (ff), Celeb-DF-domain data (cdf), and standalone sources such as internet-style or whole-image synthesis methods.

No single detector had the highest accuracy in every source group. CLIP obtained 0.940 accuracy on the FaceForensics++ group, but its real-image accuracy on the Celeb-DF group was 0.004. GenD obtained 0.732 accuracy on the standalone group and higher real-image accuracy than CLIP in the Celeb-DF and standalone groups. NPR and UnivFD were not selected as standalone image systems because their fake-image accuracy was low on several face-manipulation groups. The CLIP result is consistent with a source-dependent decision boundary: the model detects nearly all cdf fake images but also maps almost all cdf real images to the fake class. GenD changes this balance, which may follow from its restricted Layer Normalization tuning, normalized feature space, and paired real-fake training setup. NPR and UnivFD were developed primarily around generated-image cues; their high real accuracy and low fake accuracy here indicate that many face-swap and reenactment samples do not contain the artifacts these models expect.

4.2.2. Image ensemble strategies

We compared nine fusion strategies: the three single models, the unweighted mean, the submitted weighted average, majority vote, maximum fake probability, minimum fake probability, and a two-model CLIP+GenD average.

The submitted weighted ensemble obtained 0.855 accuracy on this validation set, compared with 0.825 for CLIP alone. It obtained 0.842 accuracy on the Celeb-DF group and 0.803 on the standalone group. The CLIP+GenD-only variant obtained higher FaceForensics++ accuracy, while the submitted three-model average obtained higher Celeb-DF and standalone accuracies. The fusion therefore trades some FaceForensics++ accuracy for fewer source-specific errors elsewhere. CLIP is strongly fake-biased in the aggregate results, while GenD and NPR have higher real accuracy; averaging their scores reduces the effect of either extreme. The max-fake and min-fake rules retain one model’s most extreme decision and consequently produce, respectively, high false-positive or high false-negative rates.

4.2.3. Image preprocessing experiments

The face cropper was motivated by a visible mismatch between DF40 training images and the CLEF test images. The DF40 training images are already face-aligned crops, while many CLEF images contain smaller faces in a full 256 by 256 frame. We therefore tested two preprocessing settings.

First, we applied the cropper to already cropped DF40 validation images. This was intentionally a negative control. Recropping already aligned images reduced overall accuracy from 0.837 to 0.610; the changes were -0.260 on DiT, -0.235 on fomm, and -0.185 on MidJourney. This result indicates that the cropper should not be applied to data already matching the aligned training format. Repeated detection changes the face scale and alignment even when the original crop is already suitable, creating a new mismatch rather than removing one.

Second, we created a CLEF-style simulation by padding DF40 face crops onto a larger canvas so that the face occupied about 20% of the image area. This makes direct inference see a smaller face, similar to the CLEF test images, while the recrop pipeline attempts to restore the training distribution. The recrop pipeline improved overall accuracy from 0.562 to 0.638 and increased real accuracy from 0.143 to 0.333. Accuracy changed from 0.540 to 0.733 on DiT, from 0.547 to 0.673 on fomm, from 0.567 to 0.747 on StyleGAN2, and from 0.593 to 0.400 on MidJourney. These results show that alignment is conditional on the input format: it is useful when it restores a small face to the scale used during training, but not as a universal preprocessing operation. The MidJourney decrease also indicates that resizing and interpolation can alter the artifact statistics used by a detector.

4.2.4. Image threshold calibration

The official submission requires hard labels, so threshold choice is part of the method. We calibrated the submitted ensemble on 250,802 DF40 frames using five-fold method-stratified cross-validation. We also trained a logistic-regression stacking model over the three probabilities $[p_{\text{CLIP}}, p_{\text{GenD}}, p_{\text{NPR}}]$. At threshold 0.5, the weighted ensemble obtained accuracy 0.905, F1 0.948, balanced accuracy 0.884, and AUC 0.954. Its balanced-accuracy-optimized threshold was 0.55, with accuracy 0.866 and F1 0.925. Stacking at threshold 0.5 obtained accuracy 0.969 and F1 0.984 but balanced accuracy 0.781 and AUC 0.951; its balanced-accuracy-optimized threshold was 0.95, with balanced accuracy 0.881. The learned normalized coefficients were close to the hand-picked weights, approximately 0.42 for CLIP, 0.38 for GenD, and 0.21 for NPR. Because the CLEF class balance was unknown during submission, we retained threshold 0.5 for the final image output. The similar coefficients explain why stacking did not improve ranking over the weighted average. Threshold optimization produced high accuracy and F1 on the fake-heavy validation records but lower balanced accuracy, so selecting that threshold would encode an assumption about the blind-set class ratio.

4.2.5. Audio local evaluation

The audio ViT-Tiny model was evaluated on a held-out local test split of 14,737 labeled samples, containing 1,242 real and 13,495 fake examples. It obtained AUC 96.65%, EER 10.1% at threshold 0.953, and min-tDCF 0.0274 at threshold 0.998. The AUC indicates that the model ranks LJSpeech and WaveFake samples separately on this split. However, the EER threshold of 0.953 and min-tDCF threshold of 0.998 show that the score scale is shifted toward the fake class. This is consistent with the 91% fake proportion and means that local AUC alone does not determine the hard-label behavior on a differently balanced test set.

4.2.6. Official evaluation

The organizers returned the official image scores in Table 1 and audio scores in Table 2. Each subset score is an accuracy value, and the final score is the official weighted aggregate. The submitted image system obtained a final score of 0.6038, and the submitted audio system obtained a final score of 0.5595.

4.3. Discussion and Generalization

The image experiments show two separate forms of distribution shift. First, the source-level results differ even when the semantic label is unchanged: CLIP classifies FaceForensics++ real images correctly but fails on most Celeb-DF real images. This suggests that real-image diversity is as important as

Table 1

Official image detection results.

Evaluation subset	Score
Baseline Deepfakes	0.7928
Organizers Real Data	0.6241
Organizers Real GT Data	0.4301
Participant Deepfake Data	0.7476
Participant Deepfake Data Weighted	0.7252
Final Score	0.6038

Table 2

Official audio detection results.

Evaluation subset	Score
Baseline Deepfakes	0.4182
Organizers Real Data	0.5396
Organizers Real GT Data	0.6271
Participant Deepfake Data	0.5377
Participant Deepfake Data Weighted	0.5323
Final Score	0.5595

fake-method diversity when constructing training and validation data. Second, the padded-image experiment shows a geometric shift caused by face scale. Face alignment partially corrects this shift, while the negative control shows that the same transformation damages inputs that are already aligned.

The image ensemble addresses model-specific errors but does not remove them. Its validation accuracy increase from 0.825 to 0.855 comes mainly from combining CLIP’s fake recall with the higher real accuracy of GenD and NPR. The official subset scores follow the same general pattern: accuracy is 0.7928 on organizer baseline deepfakes and 0.7476 on participant deepfakes, but only 0.4301 on the secret verified real subset. The lower real score indicates a remaining false-positive problem under an unseen real-image distribution. Possible contributors are real-source shortcuts learned by the image models and interpolation artifacts introduced by face cropping; the available subset scores do not isolate their individual effects.

The audio results show a larger difference between local and blind evaluation. The local AUC of 96.65% was obtained with one real-speech source and WaveFake-generated speech. In the official evaluation, accuracy was 0.4182 on baseline deepfakes, 0.5323 on realism-weighted participant deepfakes, and 0.6271 on the secret real subset. This pattern indicates that the submitted model missed many unseen fake samples rather than transferring the local LJSpeech-versus-WaveFake separation. Both audio classifiers use the same LFCC input, so score averaging changes the classifier architecture but does not add a different acoustic representation.

Three lessons follow from these experiments. Training and validation sets should contain diverse real sources as well as diverse generators; otherwise source identity can become an easier cue than authenticity. Preprocessing should be selected according to the input format instead of being applied uniformly. Finally, threshold-free local metrics should be accompanied by class-balanced and cross-source evaluation when the submission requires hard labels. For audio, future work should add multi-speaker real data, unseen vocoders, channel and codec augmentation, and a dedicated waveform or self-supervised anti-spoofing model. For images, source-balanced real data and calibration on multiple external datasets are direct targets for reducing false positives.

5. Conclusions

We submitted image and audio deepfake detection systems to ImageCLEF 2026. The image system uses conditional face alignment followed by weighted averaging of CLIP, GenD, and NPR outputs, and obtained an official final score of 0.6038. The audio system uses LFCC features with ViT-Tiny and ResNet18 score averaging, and obtained an official final score of 0.5595. Development experiments showed that the image models had different class biases and that alignment helped only when the input face scale differed from the training format. The official image score on the secret real subset and the difference between local and official audio results show that generalization was not resolved by model averaging alone. More diverse real data, generator coverage, cross-source validation, and modality-specific augmentation are needed to reduce these errors.

Acknowledgments

This work is supported by the National Social Science Foundation of China (24BYY080).

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI Codex for language editing, literature and citation organization, peer-review response support, and $\LaTeX$ and TikZ formatting; all outputs were verified and revised by the authors, who take full responsibility for the publication’s content.

References

- [1] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, B. Wu, Deepfakebench: A comprehensive benchmark of deepfake detection, in: *Advances in Neural Information Processing Systems*, volume 36, Curran Associates, Inc., 2023, pp. 4534–4565. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets_and_Benchmarks.pdf.
- [2] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, L. Yuan, C. Wang, S. Ding, Y. Wu, Df40: Toward next-generation deepfake detection, *arXiv preprint arXiv:2406.13495* (2024).
- [3] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, H. Delgado, ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection, *arXiv preprint arXiv:2109.00537* (2021). URL: <https://arxiv.org/abs/2109.00537>.
- [4] B. Ionescu, H. Müller, D.-C. Stanciu, et al., Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.
- [5] D.-C. Stanciu, A. Radu, R.-G. Bolborici, M. Negru, A.-F. Ene, V.-M. Vasilescu, A.-A. Nicolae, B. Ionescu, L.-D. Ștefan, M.-G. Constantin, M. Dogariu, A.-G. Andrei, Overview of imageclef 2026 deepfake task: Multimodal detection and generation of deepfakes, in: *CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026*.
- [6] ImageCLEF, Imageclef 2026 deepfake detection and generation task, <https://www.imageclef.org/2026/deepfake-detection-and-generation>, 2026. Accessed: 2026-05-17.
- [7] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.
- [8] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3207–3216.
- [9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language

- supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
- [10] U. Ojha, Y. Li, Y. J. Lee, Towards universal fake image detectors that generalize across generative models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24480–24489.
 - [11] C. Tan, H. Liu, Y. Zhao, S. Wei, G. Gu, P. Liu, Y. Wei, Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 28130–28139.
 - [12] A. Yermakov, J. Cech, J. Matas, M. Fritz, Deepfake detection that generalizes across benchmarks, arXiv preprint arXiv:2508.06248 (2025).
 - [13] J. Frank, L. Schönherr, Wavefake: A data set to facilitate audio deepfake detection, in: Advances in Neural Information Processing Systems Datasets and Benchmarks Track, 2021. URL: <https://arxiv.org/pdf/2111.02813>.
 - [14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
 - [15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021. URL: <https://openreview.net/forum?id=YicbFdNTTy>.
 - [16] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, N. W. D. Evans, AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6367–6371. doi:10.1109/ICASSP43922.2022.9747766.
 - [17] K. Ito, L. Johnson, The lj speech dataset, <https://keithito.com/LJ-Speech-Dataset/>, 2017.