Machine Intelligence Research
www.mi-research.net

23(3), June 2026, 501-527
DOI: 10.1007/s11633-026-1637-x

The Competition of Fairness in AI-generated Face Detection: Methods and Results

Shu Hu[1†*] Li Lin[1†] Shail Desai[1†] Aditya Pawar[1†] Guangyu Lin[1†]
Xin Wang[2†] Daniel S. Schiff[3†] Sachi Nandan Mohanty[4†] Ryan Ofman[5†]
Narcis Bejtic[6†] Jon Gillham[6†] Wenbin Zhang[7†] Baoyuan Wu[8†]
Cristian Canton[9†] Xiaoming Liu[10†] Luisa Verdoliva[11†] Siwei Lyu[12†]
Yongwei Tang[13] Zhiqiang Wu[13] Jiawen Seow[13] Zara Alaverdyan[14]
Anne-Flore Baron[14] Simon Bozonnet[14] Martins Bruveris[14] Jochem Gietema[14]
Lucia Innocenti[14] Lisa Ivanova[14] Olivier Koch[14] Harry Ni[14]
Arthur Pajot[14] Romain Sabathe[14] Fengming Gu[15] Xingming Long[15]
Jie Zhang[15] Wenqing Ge[16] Xiangkui Cao[15] Yuecong Min[15]
Yingjie Liu[16] Zonghui Guo[16] Shiguang Shan[15] Jinhee Park[17,18]
Minjun Kim[17] Ahyeon Park[17] Guisik Kim[18] Taewoo Kim[18]
YoungJoon Yoo[17] Junseok Kwon[17] Zhaoda Li[19] Mengyun Tang[19]
Leyang Huang[20] Bogdan Dura[21] Sebastian Balmuş[21]
Fang-Yi Su[22] Tsung-Hua Lee[22] Ting-Wan Kao[22]

[1]School of Applied and Creative Computing, Purdue University, West Lafayette 47907, USA
[2]Department of Epidemiology & Biostatistics, University at Albany, State University of New York (SUNY), Albany 12222, USA
[3]Department of Political Science, Purdue University, West Lafayette 47907, USA
[4]School of Computer Science and Engineering, VIT-AP University, Andhra Pradesh, Amaravati 522241, India
[5]Deep Media AI Inc., Oakland 94607, USA
[6]Originality.AI Inc., Ontario L9Y 2L6, Canada
[7]Michigan Technological University, Houghton 49931, USA
[8]School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China
[9]Barcelona Supercomputing Center, Barcelona 08034, Spain
[10]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill 27599, USA
[11]Department of Electrical Engineering and Information Technology, University of Naples Federico II, Naples 80125, Italy
[12]Department of Computer Science and Engineering, University at Buffalo, SUNY, Buffalo 14068, USA
[13]Ant International, Shanghai 200120, China
[14]Entrust Corp., Shakopee 55379, USA
[15]State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[16]Ocean University of China, Qingdao 266005, China
[17]Chung-Ang University, Dongjak-gu 06974, Republic of Korea
[18]Korea Electronics Technology Institute, Gyeonggi-do 13509, Republic of Korea
[19]Independent Researcher, Beijing 100042, China
[20]The University of Hong Kong, Hong Kong 999077, China

News&Views
Manuscript received on January 13, 2026; accepted on January 26, 2026; published online on April 13, 2026
Recommended by Associate Editor Deng-Ping Fan
Colored figures are available in the online version at https://link.springer.com/journal/11633
* Corresponding Author: Shu Hu (E-mail: hu968@purdue.edu)
† Competition Organizers.

[21] National Institute of Research and Development in Informatics - ICI Bucharest, Bucharest 011455, Romania
[22] Department of Biomedical Informatics, Harvard Medical School, Harvard University, Boston 02115, USA

Abstract: AI-generated facial synthesis (DeepFakes) has rapidly progressed through advanced generative models, producing realistic manipulated media that is increasingly difficult to distinguish from authentic content. Misuse of these technologies threatens public trust and democratic stability, particularly in sensitive contexts. Beyond detection accuracy, responsible forensic analysis demands fair and ethical deployment, yet recent studies reveal demographic performance disparities that existing methods have not adequately addressed, and fairness approaches often fail to generalize under distribution shifts. To address this issue, we organized the first competition focused on fairness in AI-generated face detection at NeurIPS 2025, connecting fairness research with real-world DeepFake challenges. The competition attracted substantial global participation, including more than 64 registered teams and 158 participants from 63 organizations across 20 countries, with 11 teams surpassing the baseline. Analysis of the submitted methods reveals that the most effective approaches combined data-centric design, robust representation learning, and model-level diversification rather than relying on fairness constraints alone. In particular, the top-ranked solution achieved strong fairness improvements by integrating careful data curation, a mixture-of-experts architecture, and test-time augmentation, demonstrating that fairness generalization can be improved without explicitly optimizing demographic-specific losses. Other competitive methods explored complementary directions, including foundation-model-based feature extraction, dual-branch fusion of global and local cues, ensemble learning, and post hoc calibration, each exposing distinct trade-offs among fairness, utility, and deployability. Our findings highlight that fairness metrics can be significantly improved through strategic system design, but also reveal limitations of current evaluation protocols and the risk of trivial solutions under fixed thresholds. Overall, this competition provides concrete empirical evidence and methodological insights for building more fair, robust, and trustworthy DeepFake detection systems, and offers guidance for future benchmarks, evaluation metrics, and responsible deployment practices. The competition website is available at: https://sites.google.com/view/aifacedetection/.

Keywords: AI-generated face, DeepFake detection, fairness, competition, cybersecurity.

Citation: S. Hu, L. Lin, S. Desai, A. Pawar, G. Lin, X. Wang, Daniel S. Schiff, S. N. Mohanty, R. Ofman, N. Bejtic, J. Gillham, W. Zhang, B. Wu, C. Canton, X. Liu, L. Verdoliva, S. Lyu, Y. Tang, Z. Wu, J. Seow, Z. Alaverdyan, A. Baron, S. Bozonnet, M. Bruveris, J. Gietema, L. Innocenti, L. Ivanova, O. Koch, H. Ni, A. Pajot, R. Sabathe, F. Gu, X. Long, J. Zhang, W. Ge, X. Cao, Y. Min, Y. Liu, Z. Guo, S. Shan, J. Park, M. Kim, A. Park, G. Kim, T. Kim, Y. Yoo, J. Kwon, Z. Li, M. Tang, L. Huang, B. Dura, S. Balmuş, F. Y. Su, T. H. Lee, T. W. Kao. The competition of fairness in AI-generated face detection: Methods and results. *Machine Intelligence Research*, vol.23, no.3, pp.501–527, 2026. http://doi.org/10.1007/s11633-026-1637-x

1 Introduction

AI-generated face (known as DeepFake) technology, powered by advanced deep neural network (DNN)-based generative AI methods such as variational autoencoders[1], generative adversarial networks (GANs)[2], and diffusion models (DM)[3], produces hyper-realistic images and videos by swapping a target individual with the faces of another without their consent. These technologies have advanced to the point where they can accurately replicate human features and expressions, making it increasingly difficult for the average person to distinguish between authentic and manipulated content. The potential for misuse of AI faces is especially alarming in sensitive areas, such as the 2024 US Presidential Election[4–6] or when they involve public figures such as Taylor Swift[7, 8]. DeepFakes′ ability to create convincing fake content that can manipulate human perception poses a profound threat to the integrity of democratic processes and individual reputations. Therefore, detecting DeepFakes is essential not only to combat misinformation but to maintain society′s trust in the information ecosystem.

Combating DeepFake technology requires a comprehensive strategy that extends far beyond the realm of mere detection, emphasizing the responsible design, development, and deployment of generative AI technologies. This field, also known as responsible forensics, focuses on applying forensic science to digital content in an ethically responsible manner, ensuring that actions to identify and mitigate DeepFakes meet high ethical standards and respect for human rights. At the heart of responsible forensics lies the commitment to fairness, especially important in the context of generative AI. It is crucial for detection tools to be crafted and used in ways that prevent unintentional bias against certain individuals or groups, thus upholding justice and equality in the digital realm. Recent studies[9–12] have highlighted biases in DeepFake detection, such as higher accuracy for lighter-skinned DeepFakes (see Fig. 1(b)), but available solutions have been less forthcoming about the nature of these biases[13, 14]. Unfairness in DeepFake detection can lead to disproportionate harms, where systems intended to prevent fraud, sexual exploitation, or political interference instead place greater burdens on minority individuals or underrepresented communities. For example, higher false positive rates for certain demographic groups may result in increased surveillance, wrongful content removal, or unjust accusations, while uneven performance across regions may weaken protections for minority countries against targeted disinformation campaigns. Such disparities risk amplifying existing social and political inequalities rather than mitigating the harms of AI-generated fake

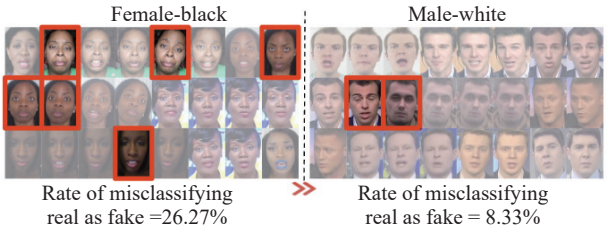


Fig. 1 Fairness challenge in DeepFake detection. The red boxes highlight the wrong predictions. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

media.

Although the machine learning community has proposed numerous fairness algorithms to help alleviate these biases[15], their application in DeepFake detection remains limited. Moreover, naively integrating fairness techniques into DeepFake detectors, without strategic implementation, often fails to ensure robust fairness under inevitable distribution shifts – a critical issue given the rapid evolution of generative AI models. For example, these models frequently produce synthetic content that deviates significantly from previously seen data, undermining the fairness generalization.

To this end, we organized the first competition focused on fairness in AI-generated face detection at the 39th Annual Conference on Neural Information Processing Systems (NeurIPS 2025), aiming to connect developments in fair machine learning with real-world challenges in computer vision. The competition is expected to catalyze new research directions, foster interdisciplinary collaboration, and encourage the design of fairer AI-generated face detection systems. Its broader impact lies in promoting the responsible deployment of AI technologies, mitigating algorithmic bias, and increasing public trust in media authenticity systems.

The objective of this competition is multi-fold: to raise awareness about the fairness challenges and opportunities in AI-generated media and its security, and to foster discussions that could lead to novel solutions and guidelines for responsible design, development, and deployment. Our competition attracted substantial global interest. We received registrations from more than 64 teams, involving 158 participants representing 63 organizations across 20 countries. Ultimately, 11 teams surpassed the provided baseline performance and were included in the leaderboard.

In this paper, we present a comprehensive summary of the competition, including its preparation and results. The remainder of the paper is organized as follows: Section 2 provides a detailed overview of the competition design and structure. Section 3 analyzes the submitted solutions, compares their performance, and highlights key findings. Section 4 briefly describes the methods proposed by teams whose submission outperformed the provided baseline. Finally, we conclude this paper in Section 5.

2 AI-face competition

This section demonstrates the significant specifics of the competition.

2.1 Datasets

2.1.1 Training set

We provide the first million-scale demographically annotated AI-Face dataset[16] to all participants as the primary training and validation sets for model development. The dataset comprises a total of 1 245 660 AI-generated face images produced by 37 different generation methods, along with 400 885 real face images. Each image is annotated with demographic attributes, including gender (e.g., male, female), age group (e.g., child, youth, adult, middle-aged adult, senior), and skin tone (based on a 10-shade scale), all of which were inferred using automated annotation tools developed by Lin et al.[16]

2.1.2 Test set

To rigorously evaluate the generalization capability and fairness robustness of submitted detectors, we curate a challenging and diverse test set consisting of 180 094 face images in total, with fully balanced real and fake samples, ensuring that performance is not biased by class imbalance.

The real face subset is sourced from the LAION-Face dataset[17], which provides large-scale, in-the-wild, high-diversity human faces reflecting real-world demographic variation. The fake subset is drawn from two sources to capture both real-world AI-generated manipulations and emerging diffusion-based synthesis threats. 73 463 images are extracted as video frames from the political DeepFakes incident database (PDID) dataset[18], the first real-world political DeepFake dataset. We include PDID due to its high societal relevance. Its inclusion underscores our emphasis on evaluating fairness in realistic high-stake deployment scenarios, where biased detection outcomes can directly influence public discourse and trust.

Additionally, 16 584 images are generated using a student diffusion model distilled from stable diffusion v1.5[19] through the improved distribution matching distillation (DMD2) method[20]. This model trained by ourselves using the LAION-Face dataset differs from the original generator. It was deliberately introduced to prevent participants from overfitting to synthetic data distributions derived directly from stable diffusion v1.5, safeguarding against potential training-set contamination or shortcut learning. By using a student model instead of the original off-the-shelf generator, we further ensure that the synthesized samples maintain realism while remaining out-of-distribution with respect to common training data, creat-

ing a stricter test condition.

Finally, demographic annotations, including gender, skin tone, and intersectional labels (defined as the combination of gender and skin tone groups) for all test samples are automatically generated using the annotation pipeline from AI-Face[16], ensuring consistency with established fairness benchmarking protocols in [16].

2.2 Task

In our competition, the AI-Face training set represents laboratory-generated AI faces, as most of these images were created for research purposes, such as designing novel detectors or benchmarking existing ones. However, relying solely on this dataset limits the ability to model real-world AI faces. Real-world AI faces are often generated by unknown actors using either single or mixed generative models, resulting in forgery types that may not have been observed during training. Additionally, the resolution and quality of faces encountered in the test set may differ significantly from those in the training set, further challenging the generalization capabilities of the developed models.

In this case, a detector trained on the provided training set, even if designed with fairness constraints in mind, may fail to maintain its fairness properties when deployed in unseen scenarios. Although this presents a significant challenge, it has not been solved sufficiently[14]. We are currently in an era of rapid advancements in generative AI, with the continual release of powerful models such as Sora 2[21]. These models can be readily misused by AI face makers to create realistic AI-generated faces, exacerbating societal risks. The dynamic nature of generative technologies creates a persistent cat-and-mouse game between AI face generation and detection, highlighting the urgent need for fairness solutions that generalize effectively to novel threats.

To this end, we introduce a significant task in our competition: Improving fairness generalization in AI face detection. This task aims to encourage participants to develop more advanced, generalizable fair AI face detectors, with the goal of creating potential solutions to mitigate these emerging challenges.

2.3 Metrics

Following the recent fairness benchmark work[16] for detecting AI-generated faces, we use four fairness metrics commonly used in the fairness community[15, 22–25] and five more widely used utility metrics[26–29]. For fairness metrics, we consider demographic parity (F_{DP})[15, 22], max equalized odds (F_{MEO})[24], equal odds (F_{EO})[23], and overall accuracy equality (F_{OAE})[24] for evaluating (e.g., individuals of a specific gender and simultaneously a specific skin tone) fairness. In experiments, the intersectional groups are female-light (F-L), female-medium (F-M), female-dark (F-D), male-light (M-L), male-medium (M-M), and male-dark (M-D), where we group 10 categories of skin tones into light (Tones 1–3), medium (Tones 4–6), and dark (Tones 7–10) for simplicity according to [30]. For utility metrics, we employ the area under the ROC curve (AUC), accuracy (ACC), average precision (AP), equal error rate (EER), and false positive rate (FPR). During evaluation, we use 0.5 as the threshold to calculate most of the metrics.

To compare detectors' performance clearly and fairly, we define the average fairness rank (Avg-F_R), which ranks each detector on each fairness metric and averages these ranks. Specifically, the Avg-F_R for detector m can be defined as follows:

$$\text{Avg-}F_R := \frac{1}{4} \sum_{f \in \{F_{DP}, F_{MEO}, F_{EO}, F_{OAE}\}} R_{m,f} \quad (1)$$

where $R_{m,f}$ is the rank of detector m for fairness metric $f \in \{F_{DP}, F_{MEO}, F_{EO}, F_{OAE}\}$.

Note that we only consider detectors that demonstrate superior AUC performance compared to the provided baselines. If multiple participants achieve the same average fairness ranks (Avg-F_R), we will compare their primary AUC performance to identify the best detector. If detectors also have identical primary utility scores, we will proceed to compare their performance on a secondary utility metric EER and recalculate their Avg-F_R for comparison. This process will be repeated iteratively until the best detector is identified. In practice, we are able to determine the competition winners without needing to exhaust all available utility metrics. The AUC score alone provides sufficient solutions to rank all participating teams.

2.4 Baselines, code, and EULA

As demonstrated in a recent fairness benchmark study[16] on AI-generated faces, there are 12 baseline models that can be used. The code and checkpoints for these baseline models can be found in the official GitHub[1]. However, we use the Xception model[31] as the baseline model for our competition due to its simplicity. Preliminary results on the AI-Face dataset are also presented in [16]. The data-loading tools necessary for handling the datasets are provided within the same repository. To access the corresponding demographic annotations of the datasets, participants are required to download and sign the end-user license agreement (EULA). The signed EULA must then be uploaded via the provided Google Form, along with the required participant information. Upon approval, the annotation download link will be sent to the requester.

[1] https://github.com/Purdue-M2/AI-Face-FairnessBench

3 Competition results

Between June 17, 2025 and October 2, 2025, we received registrations from 64 teams comprising 158 participants, representing 63 organizations across 20 countries worldwide. Most of the teams are from academia. Finally, 11 teams achieved performance surpassing the provided baseline model on the test set. Their results and rankings are presented in Table 1.

As shown in Table 1, the top two teams achieved excellent fairness performance, with most fairness disparity values close to zero. However, their overall accuracy was approximately 50% and the false positive rate reached 100%. Our inspection of their submissions shows that their prediction scores are consistently above 0.5 with very low variance. This strategy is effective in the context of our competition, where fairness is weighted more heavily than utility, as it results in all real samples being classified as fake under the predefined threshold of 0.5. However, such behavior poses challenges for real-world deployment, particularly in applications that require binary outputs for general users. Additionally, it is evident that most teams achieved significantly higher AUC scores than the baseline, indicating strong improvements.

4 Competition methods from the teams

4.1 Ant International

Members: Yongwei Tang (mail to: tangyongwei.tyw@ant-intl.com), Zhiqiang Wu (mail to: hanli.wzq@ant-intl.com), Jiawen Seow (mail to: jiawen.seow@ant-intl.com).

4.1.1 General method description

Our proposed methodology is built on three complementary pillars: 1) Quality-driven training data assessment and selection, 2) adaptive multi-feature fusion via a mixture of experts, and 3) robustness enhancement through test-time augmentation (TTA). Fig. 2 outlines the overview of our model architecture.

Quality-driven training data assessment and selection. For quality training data assessment and selection, we first performed a detailed analysis of the training set by inspecting data distributions from both intersectional-subgroup and class-balance perspectives. We then evaluated how these distributional properties affect model behavior through controlled ablation studies using a fixed architecture, allowing us to isolate the impact of dataset composition on performance and robustness. Empirically, our ablation studies showed that training on the fully pooled dataset leads to overfitting and degraded generalization. To improve robustness, we pruned the training corpus to increase both data purity, reducing distribution mismatch and noisy training signals, and data potency, retaining samples that provide strong supervisory signal. The ablation results consistently indicated that excluding the standalone GAN and DM datasets yields more stable optimization and stronger overall performance, achieving an AUC of 0.821 compared with 0.497 when training on the full pooled dataset. After refining the dataset to better match the target distribution, we further expanded training diversity by incorporating DF40[32], an external dataset containing 40 generation methods with source images from FF++[33], CelebA[34], Celeb-DF[35], and FFHQ[36]. Because DF40 closely aligns with the final AI-Face distribution, it serves as a high-quality supplement that broadens exposure to manipulation types and improves generalization to unseen generation techniques, increasing the AUC to 0.898. Table 2 shows the key ablation studies for data selection.

Table 1 Results of the competition of fairness in AI face detection. We only report the teams′ performance, which is over baseline performance. The final ranking is determined by Avg-F_R and then AUC.

Team	Rank	Avg-F_R ↓ (%)	AUC ↑ (%)	Fairness metrics (%)				Utility metrics (%)			
				F_{DP}	F_{EO}	F_{MEO}	F_{OAE}	ACC	AP	EER	FPR
Ant International	1	3.00	95.46	0.00	0.00	0.00	78.10	50.00	95.14	8.73	100.00
NDF	2	3.00	64.06	0.00	0.00	0.00	78.10	50.00	56.51	41.05	100.00
Team Entrust	3	4.00	70.21	0.06	16.97	16.67	0.12	50.01	76.03	49.99	49.99
AIST-ICT	4	4.25	70.28	0.07	0.10	0.07	78.11	50.01	71.21	35.18	99.99
CAU-KETI	5	4.50	37.06	0.49	1.93	0.55	77.73	49.85	45.05	60.21	99.87
qqpprun	6	5.00	90.47	1.71	4.33	1.70	77.58	50.64	87.33	15.97	98.69
AI-Hunter	7	5.50	59.95	0.43	4.41	2.21	78.06	50.52	58.30	40.54	98.82
Noname2	8	6.25	42.41	8.59	46.13	16.67	70.57	49.70	46.69	56.46	91.89
ICI Innolabs	9	7.00	78.09	51.31	61.81	25.03	18.96	78.09	70.95	21.91	29.06
Verix AI	10	7.50	72.13	5.88	155.57	65.77	76.02	48.51	73.30	28.02	98.28
kunkun	11	8.00	72.36	44.80	280.38	87.58	30.29	67.38	75.18	28.07	39.48
Baseline	–	–	31.32	47.03	292.83	88.41	46.16	36.29	43.63	68.85	53.32

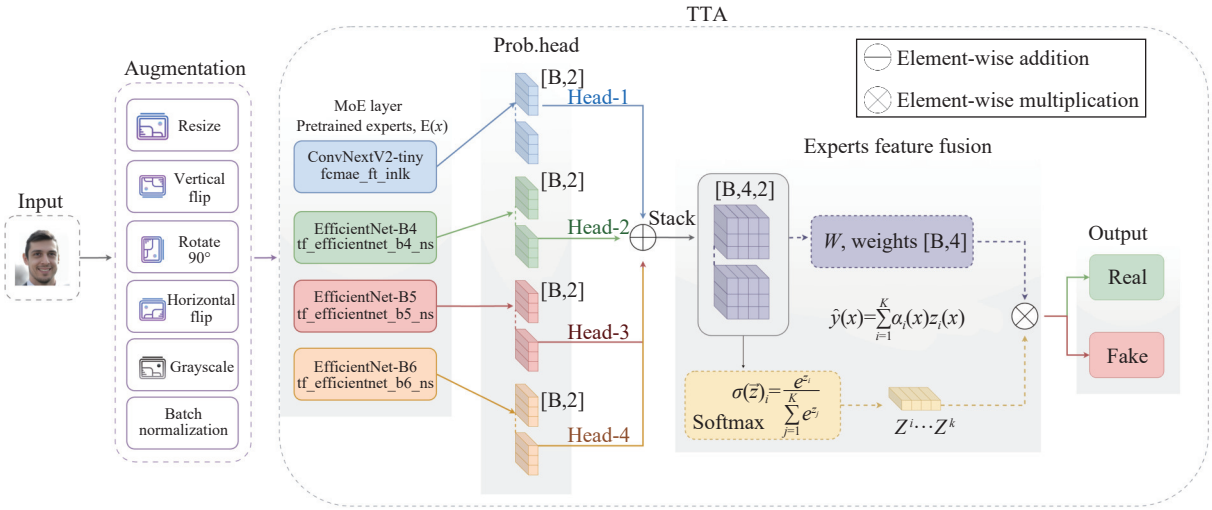


Fig. 2 Team Ant International: Model overview. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

Table 2 Key ablation studies for data selection

Datasets	AUC	Avg-F_R	EER	FPR	AP	ACC
Real + DeepFakes + GANs + DM	0.497	0.381	0.499	0.681	0.500	0.500
Real + DeepFakes	0.821	0.550	0.257	0.118	0.825	0.595
Real + DeepFakes + D40	0.898	0.211	0.182	0.982	0.910	0.508

Adaptive multi-feature fusion via a mixture of experts. We propose multi-feature fusion mixture of experts (MMF_MoE), a multi-expert fusion architecture that integrates diverse and complementary feature representations produced by multiple specialized experts. The experts are instantiated as different pretrained backbones, leveraging their distinct inductive biases. Given an input sample x, each expert E produces a prediction output. In the formulation, we denote the expert output as

$$\mathbf{z}_i(x) = E_i(x), \qquad i = 1, \cdots, K. \tag{2}$$

A dedicated gating network $g(\cdot)$ then generates softmax-normalized mixture weights for input-adaptive expert selection:

$$\boldsymbol{\alpha}(x) = \text{softmax}(g(x)), \qquad \sum_{i=1}^{K} \alpha_i(x) = 1,\ \alpha_i(x) \geq 0. \tag{3}$$

The final prediction is computed by a weighted aggregation of expert outputs:

$$\hat{\mathbf{y}}(x) = \sum_{i=1}^{K} \alpha_i(x)\, \mathbf{z}_i(x). \tag{4}$$

This adaptive softmax-gated fusion emphasizes the most informative expert(s) for each instance, rather than relying on static averaging, thereby improving robustness and generalization. It can also support fairness by limiting over-dependence on any single feature pathway that may encode group-specific artifacts. Although MMF_MoE follows a mixture-of-experts design, we do not explicitly train experts to specialize in specific demographic subgroups. Instead, we treat them as high-capacity ensemble members whose relative strengths may vary across intersectional subgroups. In the implementation, we use ConvNeXt[37] and EfficientNet[38] due to their strong baseline performance in deepfake detection. Each expert is initialized from weights pretrained on large-scale deepfake datasets and subsequently fine-tuned on their curated training set. We hypothesize that the experts capture complementary cues and may perform differently across subgroups. For example, ConvNeXt may be stronger for certain younger male groups with moderate skin tones, while EfficientNet-B4 may be stronger for some female groups. At inference time, the expert outputs are stacked and passed through the softmax gating network to produce per-expert weights, which are then used to aggregate the final detection score.

Robustness enhancement through TTA. To further boost model performance and mitigate bias, we adopt TTA, in which multiple transformed views of the same input are evaluated and then aggregated into a single prediction. We conducted an ablation study across several TTA aggregation strategies and found that max aggregation over augmented predictions consistently yields the best performance. This result aligns with our

hypothesis that max-pooling across TTA outputs acts similarly to a mixture-of-experts max-pooling strategy, by emphasizing the most reliable prediction and reducing worst-case errors. Formally, for M test-time transforms $t_m(\cdot)$, we compute

$$\hat{\mathbf{y}}_{\mathrm{TTA}}(x) = \max_{m\in\{1,\cdots,M\}} \hat{\mathbf{y}}(t_m(x)) \,. \tag{5}$$

The TTA transformations include resizing images to 512×512, horizontal and vertical flips, 90-degree rotations, grayscale, and batch normalization. Based on the analysis shown in Fig. 3, we found that the 90° rotation augmentation contributes the most to the final prediction. Table 3 outlines the ablation studies among the TTA strategies.

4.1.2 Training details

The proposed model contains approximately 188.27 M parameters and is trained using binary cross-entropy (BCE) loss. Training is performed in PyTorch on 4× NVIDIA RTX PRO 6 000 GPUs, and takes 5 hours per epoch. We optimize the network with AdamW using a learning rate of 1×10^{-4}, a batch size of 256. We train 10 epochs with a CosineAnnealingLR schedule. The computational cost is approximately 105.6 B FLOPs.

4.2 Team Entrust

Members: Zara Alaverdyan (mail to: sarah.alaverdyan@entrust.com), Anne-Flore Baron (mail to: anneflore.baron@entrust.com), Simon Bozonnet (mail to: simon.bozonnet@entrust.com), Martins Bruveris (mail to: martins.bruveris@entrust.com), Jochem Gietema (mail to: jochem.gietema@entrust.com), Lucia Innocenti (mail to: lucia.innocenti@entrust.com), Lisa Ivanova (mail to: lisa.ivanova@entrust.com), Olivier Koch (mail to: olivier.koch@entrust.com), Harry Ni (mail to: harry.ni@entrust.com), Arthur Pajot (mail to: arthur.pajot@entrust.com), Romain Sabathe (mail to: romain.sabathe@entrust.com).

4.2.1 General method description

The submission of our team unfortunately optimized for the wrong metric: Instead of optimizing the average rank of the four fairness metrics, we optimized for the average of the numerical values of the four fairness metrics. The distinction between rank and numerical average is important, as each leads to a different optimal solution. If we want to optimize for the average rank, we should use a classifier f, which assigns each image x a score $f(x) > 0.5$ above the decision threshold $\theta = 0.5$, i.e., a classifier that declares each image to be fake. This classifier will satisfy $F_{DP} = 0$, $F_{EO} = 0$ and $F_{MEO} = 0$. Thus,

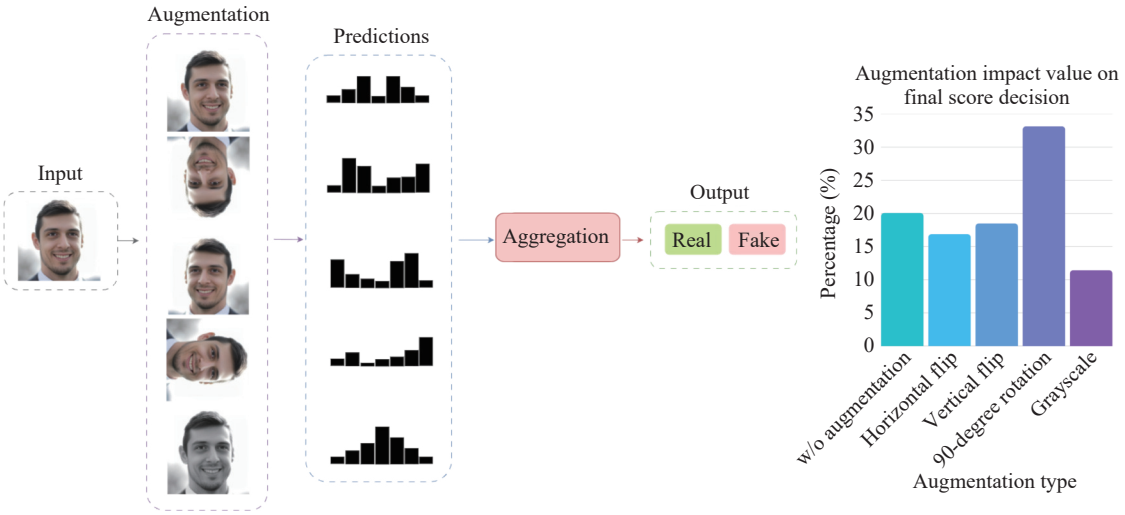


Fig. 3 Team Ant International: TTA module overview (Colored figures are available in the online version at https://link.springer.com/journal/11633)

Table 3 TTA ablation studies

Models	AUC	Avg-F_R	EER	FPR	AP	ACC
Baseline	0.313	1.186	0.688	0.533	0.436	0.363
Random	0.499	0.182	0.500	0.500	0.498	0.499
MMF_MoE (pretrained)	0.655	0.683	0.351	0.131	0.612	0.510
MMF_MoE (w/o TTA)	0.854	0.197	0.234	0.999	0.869	0.500
MMF_MoE (TTA mean)	0.898	0.211	0.182	0.982	0.910	0.508
MMF_MoE (TTA max)	0.955	0.196	0.087	1.000	0.951	0.500

this classifier is guaranteed to have (shared) rank 1 for these three metrics. Of course, this classifier will not have $F_{OAE} = 0$, as the differences in accuracy across the demographic groups will reflect the varying ratios of genuine to fake images across the groups. But having rank 1 on three out of four metrics outweighs the penalty one has to pay for a large value for the fourth metric. This is in fact the strategy adopted by the top teams. Having obtained the desired values for the fairness metrics, what about the utility metrics that are used as a tie-breaker? The AUC of such a classifier can be arbitrarily high, because we can make the scores of the classifier discriminative whilst keeping them above the decision threshold. This is possible, because all four fairness metrics only depend on the behavior of the classifier at the decision threshold $\theta = 0.5$, while the utility metric AUC is threshold-independent. In other words, the winning strategy is to train as accurate a model as possible without any need to consider bias and then to rescale the scores of the model to always lie above the decision threshold $\theta = 0.5$ that is used to measure fairness metrics.

On the other hand, if we optimize for the numerical average of the fairness metrics, as our team did, then the optimal solution is different: In the theoretical limit of an infinite dataset, it is a random classifier, assigning an image to be genuine or fake with a probability of 0.5. This classifier will achieve the perfect value of 0 for all four fairness metrics. This is because for such a classifier all relevant performance metrics will have value of 0.5:

$$\forall a \in A, \quad \mathrm{PR}_a = 0.5, \quad \mathrm{FPR}_a = 0.5,$$
$$\mathrm{FNR}_a = 0.5, \quad \mathrm{Acc}_a = 0.5$$

where A is the group set. The AUC of this classifier is 0.5, which is already higher than the AUC of the baseline Xception model, which has an AUC of 0.313 237. It should be noted that the AUC of the baseline model is lower than that of a random classifier.

In order to achieve Avg-$F_R = 0$ in practice, we would need to be able to achieve a value of 0.5 for all four performance metrics, PR, FPR, FNR, and Acc on all groups. This is theoretically possible on infinite datasets, and it is possible on finite datasets, if every group/label subset $S_{y,a} = \{(X, Y, a) : Y = y, A = a\}$ has an even number of elements.

However, the challenge test set does not have an even number of elements in each group/label subset. In fact, the challenge test set has a very unequal distribution of samples across groups.

1) Group 0 (light skin female) has only 3 fake images. The other 836 images are genuine.

2) Group 3 (light skin male) has only 6 fake images. The other 1 512 images are genuine.

The other groups have a more balanced, genuine/fake composition, see Table 4.

The low number of fakes is important for two reasons:

1) The target metric Avg-F_R is sensitive to the classification of the 3 fakes in group 0 and the 6 fakes in group 3. This is because the false positive rate, FPR, in these two groups is measured using only 3 and 6 samples, respectively. In group 0, a change in the classification of a single sample will change FNR by 33%. Similarly, in group 3, a single sample can change FNR by 16.66%.

2) Because the number of fakes in group 0 is odd, we cannot achieve the exact 50%-50% split discussed above, and hence no solution can achieve Avg-$F_R = 0$. The closest we can come is to classify 2 of the 3 images as fake and 1 as genuine. This gives FNR = 33.33% in group 0, resulting in $F_{EO} = 16.66\%$ and $F_{MEO} = 16.66\%$, thus Avg-$F_R = 8.33\%$.

In practice, there are some other group/label subsets with an odd number of samples, so the lowest value that can be achieved is Avg-$F_R = 8.370\%$. This can be obtained with the following allocation of samples, as shown in Table 5.

How does this solution compare to the solution that achieves optimal average rank? The model that assigns each image to be fake has $F_{OAE} = 0.781$ with the other fairness metrics being 0 and thus Avg-$F_R = 19.53\%$, which is higher than that of the random classifier.

We would like to note that this analysis was made

Table 4 Test set composition

Skin tone	0	1	2	3	4	5
Genuine	836	30 415	3 132	1 512	39 690	14 462
Fake	3	7 274	2 471	6	27 611	52 682

Table 5 Optimal solution to the challenge, achieving Avg-$F_R = 8.370\%$

Label	Prediction	0	1	2	3	4	5
Genuine	Genuine	418	15 207	1 566	756	19 845	7 231
	Fake	418	15 208	1 566	756	19 845	7 231
Fake	Genuine	1	3 637	1 235	3	13 805	26 341
	Fake	2	3 637	1 236	3	13 806	26 341

possible, because participants had access to ground truth test set labels during the challenge. As we discovered through organic exploration of the data, the test set provided by the challenge was not shuffled. The files were labeled consecutively from 0000001.png to 0180094.png and there is the following correspondence between the filenames and the labels: Images with filenames from 0016585.png up to and including 0106630.png are genuine and all others are fake.

For our submission to the challenge, we decided not to use ground-truth labels. Instead, we used a model-guided random sampling method to sample a random classifier. In order to implement this method, we need to train an accurate deepfake detection model and we describe our training process in Section 4.2.2.

4.2.2 Training strategy

We augmented the provided AI-Face-v2 dataset, using the following datasets: Authentica[39], DigiFakeAV[40], CelebA[34], Balanced Faces in the Wild[41], Deepfake Synthetic-20K[42], UTKFace[43]. Our final training set contained 2.4 M images (576 665 genuine and 1 835 300 fake images).

We trained a ConvNeXt-L model, pre-trained with contrastive language-image pre-training (CLIP) on LAION-2B. The model was optimized using AdamW with a learning rate of 5×10^{-6}, weight decay of 1.55×10^{-5}, epsilon of 1×10^{-8}, and momentum of 0.9. We employed a cosine annealing learning rate decaying from 10^{-5} to 5×10^{-8} over 20 epochs. We used a batch size of 128, dropout rate of 0.18 and gradient clipping with a maximum norm of 1.0.

All images in both the AI-Face-v2 dataset and external datasets underwent face cropping[44] before training and inference. We used the RetinaFace detector and enlarged the bounding box by a factor of 1.1. During training, we applied random bounding box jittering with a stretch factor of 0.05 and a translation factor of 0.05, relative to the bounding box size, to introduce spatial robustness. For validation, we used deterministic cropping with the same bounding box enlargement.

We applied moderate image augmentations to improve generalization: Horizontal flipping with probability of 0.3, color jitter with magnitude of 0.4 and probability of 0.01, grayscale conversion with probability of 0.01, Gaussian blur with probability of 0.01, and random erasing with probability of 0.01. This conservative augmentation strategy preserves the semantic content of faces while introducing sufficient variation to prevent overfitting.

To address the demographic imbalance in the training data, we implemented an attribute-based sampler that re-weights samples according to their demographic group. Specifically, we applied two-fold oversampling to dark skin tones, both males and females (groups 2 and 5), while maintaining equal sampling for other groups.

4.3 AIST-ICT

Members: Fengming Gu (mail to: gufengming18@mails.ucas.ac.cn), Xingming Long (mail to: xingming.long@vipl.ict.ac.cn), Jie Zhang (mail to: zhangjie@ict.ac.cn), Wenqing Ge (mail to: gwq17866836583@163.com), Xiangkui Cao (mail to: xiangkui.cao@vipl.ict.ac.cn), Yuecong Min (mail to: minyuecong@ict.ac.cn), Yingjie Liu (mail to: yingjieliu@stu.ouc.edu.cn), Zonghui Guo (mail to: guozonghui@ouc.edu.cn), Shiguang Shan (mail to: sgshan@ict.ac.cn).

4.3.1 General method description

As shown in Fig. 4, we introduce a fairness-oriented face forgery detection framework that consists of two stages: 1) Real face modeling and 2) fairness-aware finetuning. In Stage 1, the model is pretrained on a large-scale real-face dataset to model the distribution of genuine faces and to learn robust, general-purpose facial representations. In Stage 2, the model is further trained on a fairness-aware subset containing both real and fake samples, enabling improved forgery detection performance while reducing demographic bias.

Real face modeling. When a model pays excessive attention to non-generalizable forgery cues, such as features related to demographic attributes like ethnicity or age, it tends to rely on group-specific characteristics, which may lead to unfair or biased predictions. Therefore, effectively guiding the model to focus on universal and generalizable forgery features is crucial for improving its fairness. Inspired by FFDBackbone[45], learning facial representations from large-scale real-face data effectively enhances the model's ability to capture general forgery cues. As a result, such representations lead to improved generalization performance and better fairness.

Specifically, FFDBackbone[45] systematically analyzes

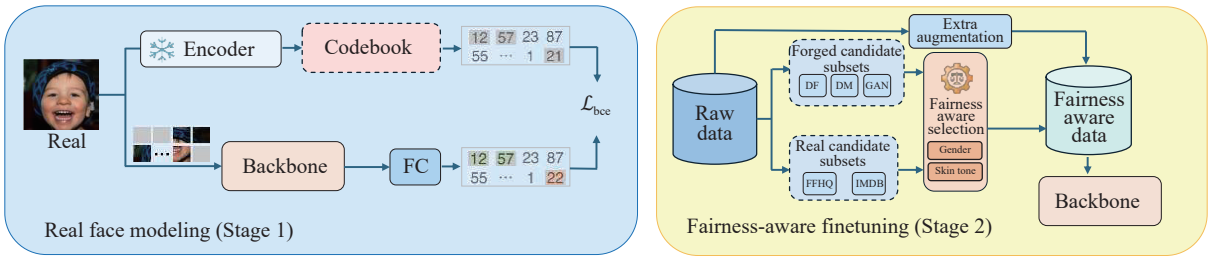


Fig. 4 Team AIST-ICT: Method overview (Colored figures are available in the online version at https://link.springer.com/journal/11633)

the critical role of the pretraining stage in facial forgery detection tasks. It employs a self-supervised learning (SSL) strategy on large-scale real-face datasets, combining contrastive learning and masked image modeling to significantly enhance the backbone's facial representation capability. Experimental results demonstrate that, compared with models pretrained on general datasets such as ImageNet, models pretrained with SSL on real-face data are better at capturing subtle forgery artifacts in local facial regions, thereby achieving stronger cross-dataset generalization. Building on these findings, we adopt FFD-Backbone as the backbone initialization and fine-tune it on our fairness-aware training subset (DeepFakes + FFHQ, augmented with Wav2Lip), thus steering the model toward universal cues and consistently reducing fairness metrics (e.g., Avg-F_R).

Fairness-aware finetuning. We observe that, despite the abundance of data spanning diverse forgery techniques[16], dataset imbalance means that additional data do not necessarily improve fairness or out-of-distribution generalization. Accordingly, we partition the training corpus into candidate subsets and select those for Stage 2 based on their measured fairness contribution. As shown in Fig. 5, an intersectional analysis over six groups defined by gender × skin tone shows that the three forgery-type subsets (DeepFakes, DMs, and GANs) have different demographic distributions; among them, the DeepFakes subset is comparatively the most balanced. When models are trained on these subsets individually, results in Table 6 show that using the DeepFakes subset as the forged data delivers the greatest fairness gains. Similarly, as shown in Table 7, within the real class, a comparison between FFHQ and IMDB reveals that FFHQ contributes more to fairness.

In addition, to compensate for the lack of audio-driven forgery methods in the training set, we randomly sample real face images from the training set as the source data and leverage Wav2Lip[46] to generate 500 000 forged faces to supplement previously unseen forgery patterns in the training set. Instead of leveraging the entire corpus, we select subsets with superior fairness properties. Specifically, we use the DeepFakes and FFHQ subsets along with Wav2Lip-generated samples. Through the careful selection of data subsets and the incorporation of synthesized samples, the resulting model achieves both strong generalization to unseen forgery methods and improved prediction fairness across demographic groups.

4.3.2 Training details

During the pretraining stage, we perform a BEiTv2-style self-supervised pretraining on the 7.5 million real-face datasets used in [45], including MSCelebV1[47], MegaFace[48], and LAION-Face[17]. This pretraining stage enhances the model's ability to learn robust and generalizable facial representations. All experiments are conduc-

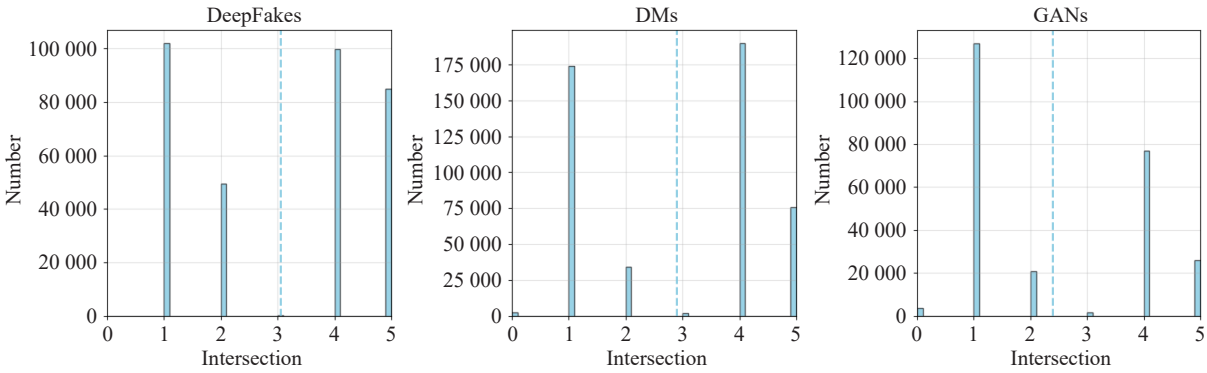


Fig. 5 Distribution of different fake training subsets

Table 6 Fairness results on validation set with different fake training subsets (lower is better)

Fake subsets	Real subsets	Avg-$F_R\downarrow$	$F_{DP}\downarrow$ (%)	$F_{EO}\downarrow$ (%)	$F_{MEO}\downarrow$ (%)	$F_{OAE}\downarrow$ (%)
DeepFakes	All	28.30	17.68	52.49	23.89	19.13
DMs	All	39.93	16.42	95.56	32.57	15.17
GANs	All	47.18	25.65	107.97	31.93	23.18

Table 7 Fairness results on validation set with different real training subsets (lower is better)

Fake subsets	Real subsets	Avg-$F_R\downarrow$	$F_{DP}\downarrow$ (%)	$F_{EO}\downarrow$ (%)	$F_{MEO}\downarrow$ (%)	$F_{OAE}\downarrow$ (%)
All	IMDB	25.14	15.22	52.42	26.83	6.08
All	FFHQ	17.80	9.29	36.20	16.62	9.09

ted on eight NVIDIA GeForce RTX 3 090 GPUs. For data preprocessing, we detect faces and extract five facial landmarks for alignment. The detected face regions are resized to a resolution of 224 × 224 and normalized with mean of 0.5 and standard deviation of 0.5. We use a batch size of 128, and optimize the model with LAMB optimizer with the cross-entropy loss. The initial learning rate is set to 5×10^{-5}, with a momentum of 0.9, and a weight decay of 0.05. A cosine annealing learning rate schedule is employed.

4.4 CAU-KETI

Members: Jinhee Park (mail to: iv4084em@cau.ac.kr), Minjun Kim (mail to: ekdh635@cau.ac.kr), Ahyeon Park (mail to: 423data@gmail.com), Guisik Kim (mail to: specialre@keti.re.kr), Taewoo Kim (mail to: kimtaewoo@keti.re.kr), YoungJoon Yoo (mail to: yjyoo3312@cau.ac.kr), Junseok Kwon (mail to: jskwon@cau.ac.kr).

4.4.1 General method description

Our approach is driven by the motivation to "correct demographic-specific biases without retraining the entire backbone". Our framework is shown in Fig. 6. It is realized through four complementary pillars:

Group-wise visual prompt tuning (VPT). Instead of full fine-tuning, we employ Shallow VPT[49] to learn specialized representations for different demographic intersections. We define $G = 6$ subgroups based on the intersection of skin tone (light, medium, dark) and gender (male, female). To mitigate data imbalance, we downsample each subgroup to 4 582 images, matching the size of the smallest intersection. We then freeze the pretrained CLIP ViT-L/14[50] backbone and the classification head of UnivFD[51].

VPT debiasing. To purify the prompts before merging, we hypothesize that demographic biases manifest as specific directions in the prompt embedding space. We implement this via a two-step process:

1) Group-centering: First, to isolate group-specific variations from the global consensus, we compute the centered prompts $\tilde{P}^{(g)}$. Let $P^{(g)}$ be the learned prompt for group $g \in \{1, 2, \cdots, G\}$, where G is the total number of groups. We subtract the average of all group prompts: $\tilde{P}^{(g)} = P^{(g)} - \frac{1}{G}\sum_{h=1}^{G} P^{(h)}$.

2) Difference-span projection: Next, we identify and remove the principal directions of demographic disparity. As illustrated in Fig. 7, we analyze the pairwise differences between centered group prompts using SVD to isolate bias components. We then explicitly project out the top-r singular vectors $V_{t,[1:r]}$ to obtain the purified token $\hat{p}_t^{(g)}$:

$$\hat{p}_t^{(g)} = \tilde{p}_t^{(g)}(I - V_{t,[1:r]}V_{t,[1:r]}^{\mathrm{T}}). \tag{6}$$

This operation filters out group-sensitive artifacts while retaining robust, shared features.

VPT merge. For merging the purified prompts, we utilize the geometric median[52] instead of the arithmetic mean. In our experiments, the simple arithmetic mean proved sensitive to outliers. However, the geometric median[52] finds the spatial center that minimizes the sum of distances to all group prompts, ensuring a more stable global prompt.

Post-hoc bias-shift calibration. Finally, to fine-tune the decision boundary for fairness, we apply a post-hoc bias shift. While the merged prompt aligns the feature space, slight discrepancies in decision thresholds may persist. We address this by adding a calibrated scalar bias b to the final logit z of the aggregated model: $\tilde{z} = z + b$.

This adjustment allows us to equalize the operating points across demographic groups, ensuring that the final decision boundary is optimized for the Avg-F_R metric. However, we highlight an inherent trade-off: While this adjustment significantly improves fairness metrics, it results in an increase in false positives.

Fig. 6 Team CAU-KETI: Method overview

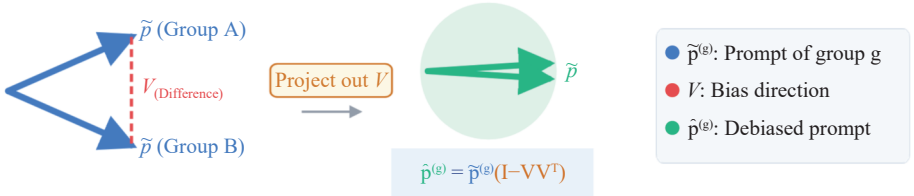


Fig. 7 Difference-span removal. We construct a difference matrix B_t by collecting all pairwise differences $\{\tilde{p}_t^{(i)} - \tilde{p}_t^{(j)}\}_{i<j}$ between centered prompts. By computing the SVD ($B_t = U_t\Sigma_t V_t^{\mathrm{T}}$), we extract the principal bias directions V_t and remove them via orthogonal projection (Equation (6)). (Colored figures are available in the online version at https://link.springer.com/journal/11633)

4.4.2 Training details

We implement our method using the CLIP ViT-L/14 architecture[50] adopted from UnivFD[51], keeping all backbone parameters frozen. For the prompt configuration, we employ shallow VPT[49] with a sequence length of 10. These tokens are initialized from a Gaussian distribution $\mathcal{N}(0, 0.02^2)$ with a dropout rate of 0.1, and crucially, their positional embeddings are set to a fixed zero tensor. All experiments are conducted on a single NVIDIA RTX 4 090 (24 GB) GPU. The training is highly efficient, involving only $\sim$ 10 k trainable parameters; it requires approximately 15 minutes per intersection-specific prompt, totaling about 1.5 hours for all six groups.

4.4.3 Quantitative analysis & discussion

Intent VS. reality. Our primary intent was to enforce demographic alignment via prompt debiasing. The results in Table 8 confirm this reality: Avg-F_R dropped significantly to 0.201 7 and F_{DP} to 0.004 9. However, we observed a trade-off where AUC decreased compared to UnivFD[51]. We attribute this to two factors: 1) The SVD projection likely removed some discriminative features along with the bias; 2) The limited capacity of shallow prompts ($\sim$ 10 k params) may be insufficient to simultaneously handle both robust generalization and strict fairness alignment. Nevertheless, we demonstrate that explicit alignment is a valid and effective direction for fairness control.

Table 8 Summarization of the performance on the Codabench test set. Arrows indicate the desired direction.

Metric	Xception	UnivFD	Ours
Avg-F_R ↓	1.186 1	0.896 2	**0.201 7**
AUC ↑	0.313 2	**0.474 7**	0.370 6
F_{DP} ↓	0.470 3	0.218 2	**0.004 9**
F_{EO} ↓	2.928 3	2.233 6	**0.019 3**

4.5 AI-Hunter

Members: Zhaoda Li (mail to: zdaiot@163.com), Mengyun Tang (mail to: merrytangmengyun@gmail.com), Leyang Huang (mail to: hleyang@connect.hku.hk).

Problem setup. Given a training dataset of size n, denoted as $\mathcal{S} = \{(X_i, D_i, Y_i)\}_{i=1}^{n}$, where X_i represents the i-th image, D_i represents the demographic attribute associated with X_i, and Y_i is the corresponding target variable (e.g., fake or real), our goal is to train a fairness-aware AI-generated face image detector that maintains high detection accuracy while ensuring balanced performance across demographic groups.

4.5.1 General method description

As shown in Fig. 8, our proposed fairness-aware dual-branch fusion network (FA-DFNet) consists of two main components: a dual-branch fusion module and a group-aware smoothing module (GASM).

1) Dual-branch fusion. The network contains a ViT branch and a ConvNeXt branch. The ViT branch captures global semantic inconsistencies through self-attention, while the ConvNeXt branch focuses on localized texture artifacts in generated images. This complementary design mitigates the blind spots of single-dimensional features by fusing global and local representations to form a more robust embedding.

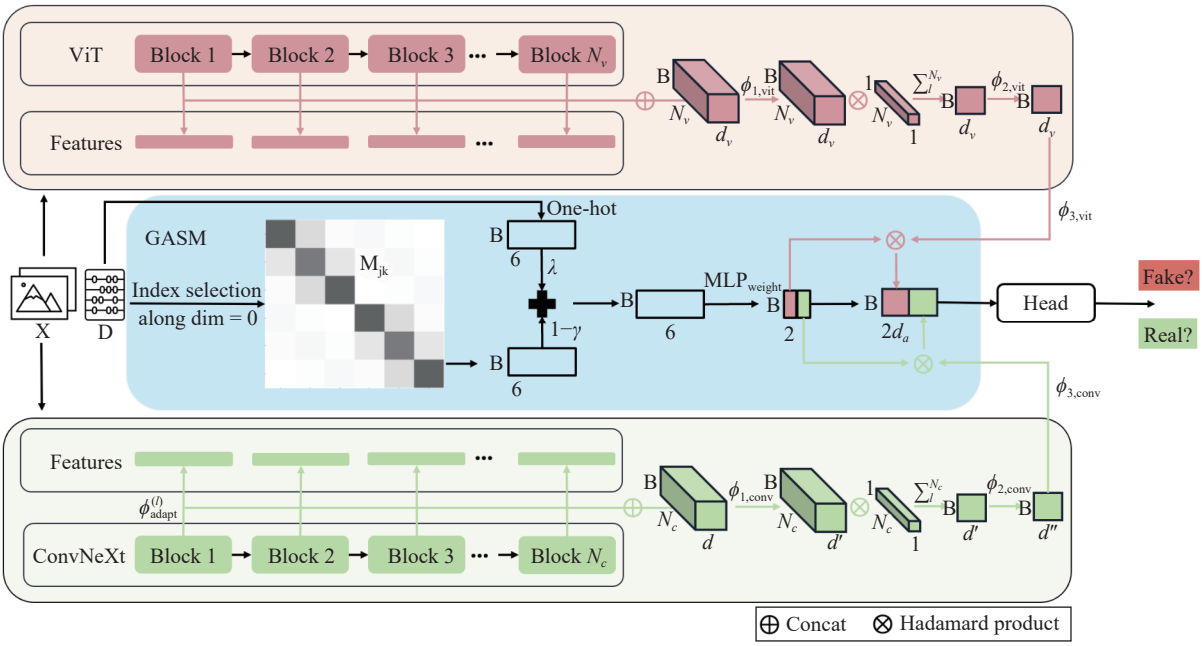


Fig. 8 Team AI-Hunter: FA-DFNet framework overview (Colored figures are available in the online version at https://link.springer.com/journal/11633)

2) GASM. The GASM enhances fairness through group probability mixing and adaptive weight generation. Group probability mixing introduces a smoothing matrix to model inter-group similarities, reducing overfitting for underrepresented groups and encouraging shared feature learning across demographics. Adaptive weight generation dynamically adjusts the fusion weights between the two branches based on the smoothed group distribution, enabling the model to adaptively balance global and local feature contributions according to demographic characteristics, thereby improving fairness.

Dual-branch. Most existing AI-generated face detection methods rely on a single type of feature – either global semantics or local textures – making them less effective against increasingly realistic generative techniques. ViTs, powered by self-attention, excel at capturing global semantic inconsistencies such as misaligned facial components or inconsistent lighting, but they struggle to perceive subtle local artifacts. In contrast, ConvNeXt, as a representative convolutional architecture, benefits from strong inductive biases that enable efficient extraction of fine-grained local textures (e.g., unnatural skin details or hair boundaries), yet it is less capable of modeling long-range dependencies. To address these limitations, we designed a dual-branch module that fuses these two complementary perspectives to form a more robust representation, enhancing the model's generalization across diverse forgery patterns. More importantly, since different demographic groups may rely on distinct feature cues, the dual-branch fusion adaptively balances this dependence, reducing inter-group performance disparities and indirectly promoting fairness.

1) ViT branch details: The ViT branch is built upon the DINOv3's[53] pretrained ViT-B/16 model. Its core idea is to fully exploit the multi-level representations from all transformer encoder layers through a hierarchical feature fusion mechanism to enhance the overall representation capacity. Given an input image $\mathbf{X} \in \mathbb{R}^{B\times 3\times H\times W}$, we feed it into the ViT backbone and extract the [CLS] token from all N_v transformer blocks. For each layer $l \in [1, N_v]$, we obtain the feature $\mathbf{cls}_{\text{vit}}^{(l)} \in \mathbb{R}^{B\times d_v}$, where d_v denotes the feature dimension of the ViT. All [CLS] tokens are then stacked to form a hierarchical feature tensor $\mathbf{G}_{\text{vit}} \in \mathbb{R}^{B\times N_v\times d_v}$. We then introduce a learnable attention weight vector $\boldsymbol{\alpha}_{\text{vit}} \in \mathbb{R}^{1\times N_v\times 1}$, which is normalized using a softmax function with a temperature parameter τ. These normalized weights are used to compute a weighted sum of the projected layer features as follows:

$$
\begin{aligned}
&\mathbf{G}_{\text{vit}} = \text{Stack}\left(\left[\mathbf{cls}_{\text{vit}}^{(1)}, \mathbf{cls}_{\text{vit}}^{(2)}, \cdots, \mathbf{cls}_{\text{vit}}^{(N_v)}\right]\right),\\
&\boldsymbol{\alpha}_{\text{vit}} = \text{Softmax}\left(\boldsymbol{\alpha}_{\text{vit}}/\tau\right),\\
&\mathbf{z}_{\text{mid,vit}} = \sum_{l=1}^{N_v} \alpha_{\text{vit},l} \cdot \phi_{1,\text{vit}}(\mathbf{G}_{\text{vit},:,l,:}),\\
&\mathbf{z}_{\text{vit}} = \phi_{2,\text{vit}}(\mathbf{z}_{\text{mid,vit}}).
\end{aligned}
$$

Specifically, $\phi_{1,\text{vit}}(\cdot) : \mathbb{R}^{d_v} \to \mathbb{R}^{d_v}$ and $\phi_{2,\text{vit}}(\cdot) : \mathbb{R}^{d_v} \to \mathbb{R}^{d_v}$ are projection networks used to fuse the layer-wise features. The final fused representation of the ViT branch is denoted as $\mathbf{z}_{\text{vit}} \in \mathbb{R}^{B\times d_v}$, which serves as the output feature vector of the ViT branch.

2) ConvNeXt branch details: The ConvNeXt branch is built upon the DINOv3's[53] pretrained ConvNeXt-Base model. Its design follows a similar principle to the ViT branch but differs in feature extraction due to the architectural characteristics of the backbone network. We extract the class tokens from all N_c stages of the ConvNeXt backbone, denoted as $\mathbf{cls}_{\text{conv}}^{(l)} \in \mathbb{R}^{B\times d_c^{(l)}}$. The key difference lies in the varying feature dimensions $d_c^{(l)}$ across stages. To address this, we introduce an independent dynamic projection adapter $\phi_{\text{adapt}}^{(l)}$ for each stage l, which projects the features into a unified dimension $d_{\max} = \max_l d_c^{(l)}$:

$$
\tilde{\mathbf{cls}}_{\text{conv}}^{(l)} = \phi_{\text{adapt}}^{(l)}(\mathbf{cls}_{\text{conv}}^{(l)}), \quad \phi_{\text{adapt}}^{(l)} : \mathbb{R}^{d_c^{(l)}} \to \mathbb{R}^{d_{\max}}.
$$

The projected class tokens from all stages are then stacked into a hierarchical tensor $\mathbf{G}_{\text{conv}} \in \mathbb{R}^{B\times N_c\times d_{\max}}$. The fusion process follows a similar formulation to that of the ViT branch:

$$
\begin{aligned}
&\mathbf{G}_{\text{conv}} = \text{Stack}\left(\left[\tilde{\mathbf{cls}}_{\text{conv}}^{(1)}, \tilde{\mathbf{cls}}_{\text{conv}}^{(2)}, \cdots, \tilde{\mathbf{cls}}_{\text{conv}}^{(N_c)}\right]\right),\\
&\boldsymbol{\alpha}_{\text{conv}} = \text{Softmax}\left(\boldsymbol{\alpha}_{\text{conv}}/\tau\right),\\
&\mathbf{z}_{\text{mid,conv}} = \sum_{l=1}^{N_c} \alpha_{\text{conv},l} \cdot \phi_{1,\text{conv}}(\mathbf{G}_{\text{conv},:,l,:}),\\
&\mathbf{z}_{\text{conv}} = \phi_{2,\text{conv}}(\mathbf{z}_{\text{mid,conv}}).
\end{aligned}
$$

Here, $\phi_{1,\text{conv}}(\cdot) : \mathbb{R}^{d_{\max}} \to \mathbb{R}^{d_c}$ and $\phi_{2,\text{conv}}(\cdot) : \mathbb{R}^{d_c} \to \mathbb{R}^{d_c}$ are projection networks.

GASM. In AI-generated face datasets, the number of samples across demographic groups is often highly imbalanced. Training directly with raw demographic labels D can lead to overfitting toward dominant groups and underfitting for minority groups, resulting in fairness issues. Moreover, demographic attributes in real-world face images are sometimes ambiguous – for example, neutral makeup may blur gender characteristics. To address these challenges, we propose GASM. The core idea is to introduce group label smoothing (GLS). This mechanism explicitly models the uncertainty of group membership while alleviating fairness problems caused by data imbalance. Conventional label smoothing replaces hard labels with soft ones to prevent overconfidence and serve as a form of regularization[54]. We extend this idea to demographic attributes by constructing a smoothing matrix $\mathbf{M} \in \mathbb{R}^{6\times 6}$, where the j-th row $\mathbf{M}_j$ represents the similarity distribution of group j with respect to other groups. For instance, the smoothed distribution of female-light (F-L, light-skinned female) assigns higher probability to female-

medium, moderate probability to female-dark and male-light, and lower probability to male-medium as well as male-dark. Specifically,

1) Group probability mixing: Given a batch of original demographic labels D (one-hot encoded as $\mathbf{O} \in \mathbb{R}^{B\times 6}$), we generate the smoothed group probability distribution $\mathbf{P}_{\text{mix}} \in \mathbb{R}^{B\times 6}$ as follows:

$$\mathbf{P}_{\text{mix}} = (1-\lambda)\cdot \mathbf{O} + \lambda \cdot \mathbf{M}[D]$$

where λ is the smoothing strength hyperparameter, and $\mathbf{M}[D]$ denotes the corresponding rows of the predefined smoothing matrix $\mathbf{M}$ selected according to the group labels D. This approach offers two main advantages: First, through an implicit knowledge transfer mechanism, samples from one group can benefit from related groups during training, effectively mitigating overfitting in minority groups. Second, it encourages the model to learn group-invariant forgery representations rather than memorizing spurious correlations tied to specific demographics, thereby improving overall fairness.

2) Adaptive weight generation: We input the smoothed group probability $\mathbf{P}_{\text{mix}} \in \mathbb{R}^{b\times 6}$ into a lightweight multilayer perceptron $\text{MLP}_{\text{weight}} : \mathbb{R}^6 \to \mathbb{R}^2$ to produce the base fusion weights:

$$\mathbf{w}_{\text{base}} = \text{MLP}_{\text{weight}}(\mathbf{P}_{\text{mix}}), \quad \mathbf{w} = \sigma(\mathbf{w}_{\text{base}}) = [w_{\text{vit}}, w_{\text{conv}}]$$

where $\sigma(\cdot)$ denotes the sigmoid activation function. The core purpose of this design is to learn group-specific feature fusion strategies. Different demographic groups may prefer distinct fusion ratios between the ViT and ConvNeXt branches: Some rely more on global semantic cues (e.g., ViT), while others are more sensitive to local texture information (e.g., ConvNeXt). The weight generation network learns this nonlinear mapping from the smoothed demographic priors $\mathbf{P}_{\text{mix}}$, and therefore overcomes the expressiveness limitations of simple linear transformations. Moreover, the entire process is fully differentiable, which enables the model to adaptively optimize the feature fusion strategy in an end-to-end manner based on demographic information.

3) Feature fusion and classification: Based on the generated fusion weights, the features from the two branches are adaptively combined as follows:

$$\begin{aligned}
\mathbf{z}'_{\text{vit}} &= w_{\text{vit}} \cdot \phi_{3,\text{vit}}(\mathbf{z}_{\text{vit}}) \\
\mathbf{z}'_{\text{conv}} &= w_{\text{conv}} \cdot \phi_{3,\text{conv}}(\mathbf{z}_{\text{conv}}) \\
\mathbf{z}_{\text{fused}} &= \text{Concat}([\mathbf{z}'_{\text{vit}}, \mathbf{z}'_{\text{conv}}]) \\
\hat{Y} &= \text{Head}(\mathbf{z}_{\text{fused}})
\end{aligned}$$

where $\phi_{3,\text{vit}}(\cdot) : \mathbb{R}^{d_v} \to \mathbb{R}^{d_a}$ and $\phi_{3,\text{conv}}(\cdot) : \mathbb{R}^{d_c} \to \mathbb{R}^{d_a}$ are feature projection layers, and $\text{Head}(\cdot) : \mathbb{R}^{2d_a} \to \mathbb{R}$ is the final classification head.

This design allows the model to dynamically adjust the contribution of each branch according to the demographic characteristics, thereby improving the detection accuracy while ensuring fairness across groups.

4.5.2 Training details

For training, we fix the batch size to 128 and train for 3 epochs using the AdamW optimizer with a weight decay of 0.05. The learning rate starts at 0.000 1 and increases linearly to 0.01. Since fairness improvements mainly come from architectural design rather than loss modification, we employ a simple BCE loss. The temperature and smoothing parameters are set to $\tau = 0.1$ and $\lambda = 0.2$, respectively. The smoothing matrix $\mathbf{M}$ is defined as follows:

$$\mathbf{M} = \begin{bmatrix}
0.70 & 0.20 & 0.03 & 0.04 & 0.02 & 0.01 \\
0.15 & 0.60 & 0.15 & 0.03 & 0.05 & 0.02 \\
0.03 & 0.20 & 0.70 & 0.01 & 0.02 & 0.04 \\
0.04 & 0.02 & 0.01 & 0.70 & 0.20 & 0.03 \\
0.03 & 0.05 & 0.02 & 0.15 & 0.60 & 0.15 \\
0.01 & 0.02 & 0.04 & 0.03 & 0.20 & 0.70
\end{bmatrix}$$

$$\text{One-Hot} = \begin{bmatrix} 0.032 & 0.016 & 0.008 & 0.76 & 0.16 & 0.024 \end{bmatrix}.$$

For example, the first row corresponds to the female-light, which assigns higher probability to female-medium, moderate probability to female-dark and male-light, and lower probability to male-dark, as well as male-medium.

4.6 ICI Innolabs

Members: Bogdan Dura (mail to: bogdan.dura@ici.ro), Sebastian Balmuș (mail to: sebastian.balmus@ici.ro).

4.6.1 General method description

Our approach combines multiple facial image classifiers within an ensemble framework in which each component model is trained with a distinct objective and inductive bias. By integrating models optimized for cross-domain generalization alongside models exhibiting more stable behavior across demographic groups, the ensemble produces detection decisions that balance predictive performance with consistent error characteristics across populations. As illustrated in Fig. 9, the proposed architecture follows a parallel inference pipeline. A single facial image is simultaneously processed by three distinct models, and their individual predictions are subsequently aggregated into a final decision through a majority voting mechanism. The ensemble consists of the following components:

1) Two MobileNetV2 variants[55]: These models share the same architecture but were trained using distinct strategies. The baseline variant focused on the com-

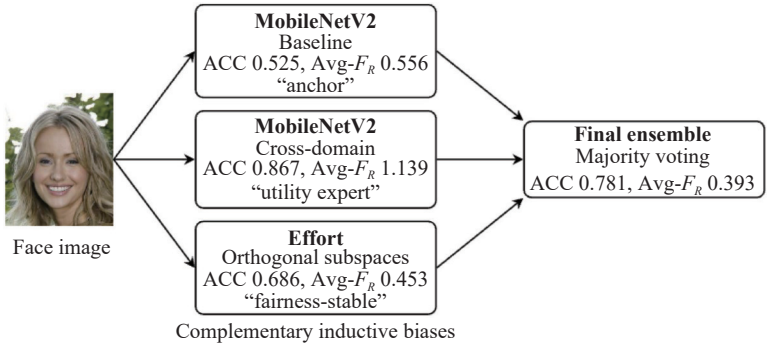


Fig. 9 Team ICI Innolabs: Architecture combining complementary inductive biases via majority voting

petition's primary dataset, while the cross-domain variant utilized an expanded dataset[56] and alternative color-space representations to improve generalization.

2) The effort model[57]: We incorporated a pre-trained effort model based on orthogonal subspace decomposition. This model separates image features into orthogonal subspaces, semantic and manipulation-related, to enhance detection under distribution shifts while maintaining stable fairness performance.

The final predictions were generated through a majority voting ensemble. This strategy was chosen to leverage the specific strengths of each model: While the cross-domain variant acted as a "utility expert" with high accuracy, the effort and baseline models provided more "fairness-stable" indicators. Our experiments demonstrated that this ensemble successfully reconciled these competing objectives, achieving the highest accuracy (ACC = 0.78) on the competition leaderboard.

4.6.2 Training details

Our training strategy was designed to enhance the robustness of the system against distribution shifts while ensuring demographic fairness through data diversity and specialized feature engineering. All experiments are conducted in a PyTorch environment on an NVIDIA RTX 4 090 GPU.

Data integration and augmentation. To mitigate overfitting and improve generalization beyond the provided AI-Face dataset, we incorporated the external "Deepfake and real images" dataset from Kaggle[56].

The first model (**MobileNetV2 baseline**) was trained exclusively on the AI-Face dataset without pre-trained weights. This variant was optimized for 30 epochs with a batch size of 256 using the Adam optimizer (learning rate $= 1 \times 10^{-4}$, weight decay $= 1 \times 10^{-4}$) and a ReduceLROnPlateau scheduler with a patience of 3 and a reduction factor of 0.15. We applied moderate data augmentations, including random horizontal flipping and color jitter (± 0.15 for brightness, contrast, and saturation; ± 0.1 for hue), without applying input normalization. Training for this configuration took approximately 18 minutes per epoch.

The second model (**cross-domain MobileNetV2**) was trained on a combined dataset consisting of the AI-Face dataset and the Deepfake and real images dataset. This model was trained from scratch for 20 epochs using the same optimization and scheduling parameters as the baseline. To encourage better generalization, images were converted to the YUV color space and subjected to stronger augmentations: enhanced color jitter (± 0.4 for brightness, contrast, and saturation; ± 0.15 for hue), Gaussian blur, and random perspective transformations. A custom normalization derived from the combined dataset was applied to all inputs. Due to the expanded data and computational overhead of the augmentations, training time increased to 29 minutes per epoch.

In addition to our trained models, we incorporated the effort model based on orthogonal subspace decomposition. We utilized a released checkpoint from prior work without additional fine-tuning to leverage its ability to separate semantic and manipulation-related features. The final pipeline employs a majority voting ensemble of these three distinct models to benefit from their complementary inductive biases.

4.6.3 Results and analysis

The evaluation of our models on the hidden test set revealed a significant trade-off between predictive utility and demographic fairness. While individual models excelled in specific areas, the ensemble approach provided the most robust and balanced performance across all metrics. Fig. 10 provides a comparative visualization of these trade-offs, showing how each model occupies a distinct region of the utility-fairness space. The performance varies considerably across the three primary architectures:

1) MobileNetV2 baseline: This model achieved a moderate accuracy of 0.525 but struggled to generalize to the hidden test distribution, resulting in a low AUC of 0.441 and a high FPR of 0.728.

2) Cross-domain MobileNetV2: By training on an expanded dataset with stronger augmentations, this variant achieved the highest individual utility metrics, including an AUC of 0.867 and an accuracy of 0.867. However, as shown in Fig. 10, this gain in utility came at the cost of demographic parity, with increased variability in fairness metrics such as F_{EO} and Avg-F_R.

3) Effort model: This model provided a more conservative trade-off, maintaining competitive fairness stabil-

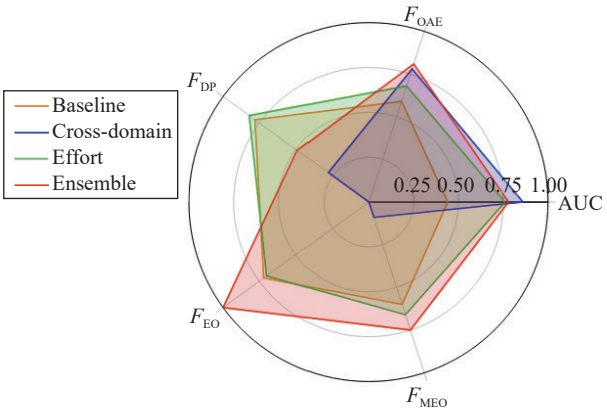


Fig. 10 Radar plot over AUC and four fairness metrics. Values are min–max normalized per metric and inverted where necessary so that larger radius always means better performance. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

ity, including the lowest F_{DP} of 0.183, at the expense of lower overall AUC (0.763) compared to the cross-domain variant.

Ensemble synthesis and competition outcome. Our final ensemble configuration, which combined these models via majority voting, successfully mitigated the weaknesses of individual components. The ensemble achieved a utility-fairness balance with an accuracy of 0.781 and an Avg-F_R of 0.393. This represents a significant improvement in demographic parity over the high-utility cross-domain model while maintaining high classification performance. Notably, our ensemble achieved the highest accuracy (ACC = 0.78) on the competition leaderboard, significantly outperforming the second-best submission (ACC = 0.56). Despite this leading accuracy, the cumulative impact of fairness penalties resulted in a final ranking of the 9th place. This outcome underscores the persistent challenge of minimizing demographic bias without sacrificing the raw predictive power required for effective AI-generated face detection.

4.7 kunkun

Members: Fang-Yi Su (mail to: fang-yi_su@hms.harvard.edu), Tsung-Hua Lee (mail to: tsunghua_lee@hms.harvard.edu), Ting-Wan Kao (mail to: tingwan_kao@hms.harvard.edu).

4.7.1 General method description

We propose the fairness aware ensemble calibration (FAEC) framework, which is a simple yet effective two stage framework designed to address fairness in deepfake detection without requiring model retraining. As illustrated in Fig. 11, the approach consists of two primary modules which are ensemble fusion for utility maximization and fairness calibration using controlled group mean interpolation (CGMI).

Stage I: Multi-model ensemble fusion. In the first stage, we maximize detection utility by aggregating predictions from five diverse architectures to leverage complementary strengths. The ensemble includes SPSL[58], Xception[31], CORE[59], F3Net[60], and EfficientNet-B4[38], which provide diversity across spatial, frequency, and reconstruction domains. For a given test image x_i, the base ensemble score $s_i^{(0)}$ is calculated via uniform averaging:

$$s_i^{(0)} = \frac{1}{M}\sum_{m=1}^{M} f_m(x_i) \tag{7}$$

where $M = 5$ represents the number of pretrained models. This strategy reduces prediction variance and ensures that no single model failure mode dominates the output.

Stage II: Fairness calibration (CGMI). The second stage applies the CGMI to align the score distributions across different demographic groups. Let a_i be

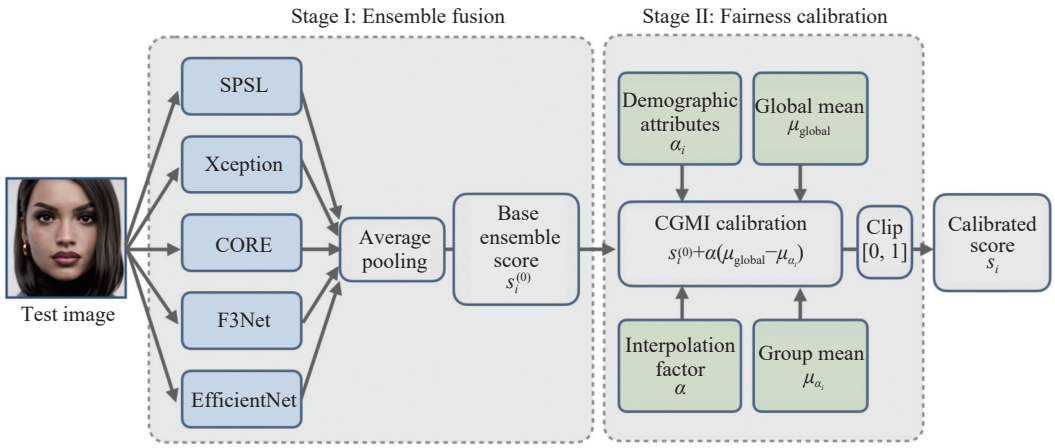


Fig. 11 Overview of the FAEC framework. Stage I aggregates predictions from diverse pre trained models to form a base score. Stage II applies CGMI using demographic attributes and statistics to produce a fairness calibrated prediction. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

the demographic attribute of sample x_i belonging to one of the eighteen intersectional groups. The final calibrated score s_i is computed as

$$s_i = \text{clip}_{[0,1]}[s_i^{(0)} + \alpha \cdot (\mu_{\text{global}} - \mu_{a_i})] \tag{8}$$

where μ_{global} is the average prediction across the test set and μ_{a_i} is the mean prediction for the specific demographic group. The interpolation factor α controls the utility fairness tradeoff.

This method satisfies three theoretically motivated properties. First, it ensures group mean convergence where group averages move toward the global mean. Second, it achieves disparity reduction by shrinking the distance between group means by a factor of $(1 - \alpha)$. Finally, it maintains rank preservation within each group to ensure that the discriminative power of the ensemble is not compromised.

4.7.2 Training strategy

Our framework is model agnostic and operates entirely in inference mode without requiring ground truth labels on the test data. The implementation uses the PyTorch 2.0 library and is executed on NVIDIA A100 GPU hardware. Images are processed at a resolution of 224 × 224 with ImageNet normalization and a batch size of 64. We employ an interpolation factor α of 0.4 based on empirical validation. This setting is chosen to reduce demographic disparity by 60% while maintaining 60% of the original group characteristics to preserve utility. The calibration step is computationally efficient as it only requires simple group level statistical adjustments after the model forward passes.

4.8 Summary

Core ideas of the proposed methods. The methods presented by the participating teams reflect diverse but complementary perspectives on fairness-aware AI-generated face detection. Ant International focused on a data-driven and architecture-driven strategy, combining careful data curation, a mixture-of-experts framework, and test-time augmentation to improve robustness and fairness simultaneously. Team Entrust approached the problem from an evaluation-aware perspective, analyzing the structure of the fairness metrics and exploiting their dependence on fixed decision thresholds to achieve optimal fairness rankings. AIST-ICT emphasized representation learning, proposing a two-stage framework that first models genuine face distributions using large-scale real-face pretraining and then performs fairness-aware fine-tuning through selective data subset construction. CAU-KETI adopted a prompt-based debiasing approach using frozen foundation models, where demographic-specific prompts were learned, purified, and merged to correct bias without retraining the backbone. AI-Hunter introduced a dual-branch architecture that fuses global semantic features and local texture cues, augmented with group-aware smoothing to encourage shared representations across demographic groups. ICI Innolabs employed an ensemble strategy, combining multiple models with complementary inductive biases to balance detection accuracy and fairness stability. Finally, kunkun primarily relied on score-level calibration by aggregating predictions from diverse detectors rather than architectural or data-centric innovations to optimize fairness metrics.

Applicable scenarios of different approaches. These methods are suitable for different deployment scenarios depending on data availability, computational constraints, and application requirements. Data-centric and mixture-of-experts approaches, such as those from Ant International and AIST-ICT, are well suited for large-scale training environments where extensive datasets and computational resources are available, and where robustness to unseen generative models is critical. Prompt-based and lightweight adaptation strategies, exemplified by CAU-KETI and kunkun, are particularly applicable in scenarios where retraining large foundation models is infeasible, such as edge deployment or rapid model adaptation. Evaluation-aware and post-processing strategies, as demonstrated by Team Entrust, may be effective in controlled benchmarking or auditing settings but are less suitable for direct deployment. Dual-branch and ensemble-based methods, proposed by AI-Hunter and ICI Innolabs, respectively, are appropriate for real-world systems that must handle heterogeneous forgery patterns and distribution shifts, especially when balancing performance across diverse demographic groups is a priority.

Advantages of the proposed methods. A key advantage of data-centric approaches is their ability to directly address bias at the source by improving training distributions, often leading to strong generalization and stable fairness gains. Model-centric designs, including mixtures of experts, dual-branch fusion, and ensemble learning, benefit from capturing complementary feature cues, which improves robustness against increasingly realistic generative techniques. Foundation-model-based approaches leverage powerful pretrained representations, enabling strong performance even with limited task-specific training. Prompt-based debiasing methods offer efficiency and interpretability, as demographic effects can be explicitly analyzed and controlled in the representation space. Post-processing and evaluation-aware strategies provide clear insights into the interaction between fairness metrics and decision rules, highlighting important limitations of current evaluation protocols and motivating more careful metric design.

Disadvantages and limitations. Despite their strengths, each class of methods also has notable limitations. Data-centric strategies depend heavily on the availability of diverse and well-curated datasets, which may not always be feasible and can introduce new biases if external data are misaligned. Complex model architectures,

such as mixtures of experts and multi-branch networks, increase computational cost and system complexity, potentially hindering scalability and reproducibility. Prompt-based and frozen-backbone approaches may sacrifice detection utility when aggressive debiasing removes discriminative features. Ensemble methods, while robust, can be difficult to interpret and maintain, especially in real-time systems. Finally, post-processing and evaluation-aware solutions risk overfitting to specific metrics or thresholds, raising concerns about robustness, transparency, and practical deployment.

5 Discussions and conclusions

Through the organization of this AI-Face competition, we learned a lot from the results. Specifically,

1) The competition indeed received wide interest around the world with 64 registered teams from 20 countries. We still receive the registered teams even after the registration deadline. To be fair, we did not include the later registered teams into the competition but this pushes us to organize the second round AI-Face competition in the near future.

2) By analyzing the results from the top teams, it motivates us to rethink how to prevent the teams from trying to achieve good fairness but lose the utility. One potential direction is to design a weighted evaluation metric that jointly accounts for utility and fairness performance. Another option is to adopt a Pareto frontier analysis to explicitly characterize the trade-offs between utility and fairness. In this setting, each team could report multiple (utility, fairness) performance pairs on the test set, obtained from different training stages, enabling the construction of trade-off curves. Based on these curves, a ranking strategy could be developed to identify teams that achieve strong and robust balance between utility and fairness. For example, teams could be compared at matched utility levels, and their average fairness performance could be used as the ranking criterion.

3) Most of the teams can achieve better AUC performance than our provided baseline, which indicates that our baseline is too simple. Therefore, this motivates us to improve the difficulty of the competition by introducing a more advanced baseline model.

4) Careful data curation prior to model training, informed by a systematic understanding of the strengths and limitations of the training data, emerged as a common strategy. In particular, Teams Ant International, AIST-ICT, and CAU-KETI analyzed the distributions of fake samples originating from different sources to guide their data preparation and training processes. Notably, Team Entrust conducted an analysis of the test set to infer both class labels and group labels, which substantially informed and improved their final model design.

5) Rather than relying on a single model, several teams adopted mixtures-of-experts architectures (in Teams Ant International, AI-Hunter, ICI Innolabs, and kunkun) and data augmentation strategies (in Teams Ant International, Team Entrust, AIST-ICT, and ICI Innolabs), which proved effective and demonstrated improved performance in the experiments.

6) By leveraging foundation models to extract more accurate and transferable feature representations, multiple teams achieved notable performance improvements. For instance, Team Entrust and CAU-KETI leveraged the latest CLIP model, while AI-Hunter employed DINOv3.

7) Several teams augmented the competition-provided dataset with external data during model training. For instance, Ant International incorporated publicly available real and DeepFake datasets, while Team ICI Innolabs leveraged an additional DeepFake dataset sourced from Kaggle. This strategy can improve the generalization of trained models.

Overall, it is encouraging to observe a diverse range of new approaches during the competition, spanning data preprocessing strategies to novel model architecture designs, aimed at improving fairness in AI-generated face detection. At the same time, the competition revealed several pressing challenges that warrant further investigation. Specifically,

1) How to prevent teams from inferring test-set labels warrants further investigation. As noted by Team Entrust, ground-truth labels in the test set could be easily identified through image file names, an issue that should be addressed in future competition designs.

2) Although participating teams demonstrated notable fairness improvements over the provided baselines, fairness remains an open problem in AI-Face detection, particularly in achieving an effective balance between utility and fairness.

3) The competition settings and evaluation metrics could be further refined to discourage trivial solutions that improve fairness at the expense of utility, thereby encouraging more practically deployable methods.

4) On the competition day, we discussed their developed approaches with top-performing teams. Some teams highlighted the use of post-processing techniques to further enhance fairness. This motivates a deeper examination of whether and how such post-processing should be permitted to ensure equitable and meaningful comparisons.

5) To stimulate more innovative and advanced solutions, future competitions could be made more challenging, for example by extending beyond facial imagery to broader visual contexts and explicitly examining trade-offs between performance on faces and other types of images.

We leave these directions for future work, with the goal of continuously improving the competition design and expanding its broader impact, ultimately advancing responsible and fair AI approaches for combating AI-gen-

erated fake media. This work also offers substantive insights that may serve as useful references for future competition organizers, both within this research area and beyond.

Acknowledgements

We thank the challenge sponsors: Deep Media AI Inc. and Originality.AI Inc. We also express our gratitude to all the participants and the organizing chairs of NeurIPS 2025 for their support in making this competition possible.

Shu Hu is supported by the USA National Science Foundation (NSF) (No. IIS-2434967) and the National Artificial Intelligence Research Resource (NAIRR) Pilot and Texas Advanced Computing Center (TACC) Lonestar6, USA. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the offcial views or policies of NSF and NAIRR Pilot.

Open Access

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] A. Vahdat, J. Kautz. NVAE: A deep hierarchical variational autoencoder. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, Article number 1650, 2020.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 8110–8119, 2020. DOI: 10.1109/CVPR42600.2020.00813.

[3] P. Dhariwal, A. Nichol. Diffusion models beat GANs on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Article number 672, 2021.

[4] AFP. Image of Biden planning military action in fatigues is fake, [Online], Available: https://factcheck.afp.com/doc.afp.com.34H74GF, 2024.

[5] Futurism. Trump's dad resurrected via AI to tell son he's a disgrace, [Online], Available: https://futurism.com/the-byte/trump-dad-ai, 2024.

[6] L. Sforza. Trump says red marks on hands may have been AI, [Online], Available: https://thehill.com/homenews/campaign/4441928-trump-says-red-marks-on-hands-may-have-been-ai/, 2024.

[7] S. M. Kelly. Explicit, AI-generated Taylor Swift images spread quickly on social media, [Online], Available: https://www.cnn.com/2024/01/25/tech/taylor-swift-ai-generated-images, 2024.

[8] T. Hsu. Fake and explicit images of Taylor swift started on 4chan, study says, [Online], Available: https://www.nytimes.com/2024/02/05/business/media/taylor-swift-ai-fake-images.html, 2024.

[9] L. Trinh, Y. Liu. An examination of fairness of AI models for deepfake detection. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, Montreal, Canada, pp. 567–574, 2021. DOI: 10.24963/ijcai.2021/79.

[10] Y. Xu, Terhörst, M. Pedersen, K. Raja. Analyzing fairness in deepfake detection with massively annotated databases. *IEEE Transactions on Technology and Society*, vol. 5, no. 1, pp. 93–106, 2024. DOI: 10.1109/TTS.2024.3365421.

[11] A. V. Nadimpalli, A. Rattani. GBDF: Gender balanced DeepFake dataset towards fair DeepFake detection. In *Proceedings of Computer Vision, and Image Processing. ICPR International Workshops and Challenges*, Montreal, Canada, pp. 320–337, 2022. DOI: 10.1007/978-3-031-37742-6_25.

[12] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, H. Malik. DeepFakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023. DOI: 10.1007/s10489-022-03766-z.

[13] Y. Ju, S. Hu, S. Jia, G. H. Chen, S. Lyu. Improving fairness in deepfake detection. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, USA, pp. 4655–4665, 2024. DOI: 10.1109/WACV57701.2024.00459.

[14] L. Lin, X. He, Y. Ju, X. Wang, F. Ding, S. Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 16815–16825, 2024. DOI: 10.1109/CVPR52733.2024.01591.

[15] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, vol. 54, no. 6, Article number 115, 2022. DOI: 10.1145/3457607.

[16] L. Lin, Santosh, M. Wu, X. Wang, S. Hu. AI-face: A million-scale demographically annotated AI-generated

face dataset and fairness benchmark. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 3503–3515, 2025. DOI: 10.1109/CVPR52734.2025.00332.

[17] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, F. Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 18697–18709, 2022. DOI: 10.1109/CVPR52688.2022.01814.

[18] C. P. Walker, D. S. Schiff, K. J. Schiff. Merging AI incidents research with political misinformation research: Introducing the political deepfakes incidents database. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, pp. 23053–23058, 2024. DOI: 10.1609/aaai.v38i21.30349.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 10684–10695, 2022. DOI: 10.1109/CVPR52688.2022.01042.

[20] T. Yin, M. Gharbi, T. Park, R. Zhang, E. Shechtman, F. Durand, B. Freeman. Improved distribution matching distillation for fast image synthesis. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Vancouver, Canada, Article number 1505, 2024.

[21] OpenAI. Sora 2, [Online], Available: https://openai.com/index/sora-2/, 2025.

[22] X. Han, J. Chi, Y. Chen, Q. Wang, H. Zhao, N. Zou, X. Hu. FFB: A fair fairness benchmark for in-processing group fairness methods. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

[23] J. Wang, X. E. Wang, Y. Liu. Understanding instance-level impact of fairness constraints. In *Proceedings of the 39th International Conference on Machine Learning*, Baltimore, USA, pp. 23114–23130, 2022.

[24] H. Wang, L. He, R. Gao, F. P. Calmon. Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 1176, 2023.

[25] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, Cambridge, USA, pp. 214–226, 2012. DOI: 10.1145/2090236.2090255.

[26] Z. Yan, Y. Luo, S. Lyu, Q. Liu, B. Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 8984–8994, 2024. DOI: 10.1109/CVPR52733.2024.00858.

[27] H. Ren, L. Lin, C. H. Liu, X. Wang, S. Hu. Improving generalization for AI-synthesized voice detection. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, Pennsylvania, USA, pp. 20165–20173, 2025. DOI: 10.1609/aaai.v39i19.34221.

[28] Z. Yan, Y. Zhao, S. Chen, M. Guo, X. Fu, T. Yao, S. Ding, Y. Wu, L. Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 12615–12625, 2025. DOI: 10.1109/CVPR52734.2025.01177.

[29] J. Cheng, Z. Yan, Y. Zhang, L. Hao, J. Ai, Q. Zou, C. Li, Z. Wang. Stacking brick by brick: Aligned feature isolation for incremental face forgery detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 13927–13936, 2025. DOI: 10.1109/CVPR52734.2025.01300.

[30] E. Monk. The monk skin tone scale. *Open Science Framework*, [Online], Available: https://osf.io/preprints/socarxiv/pdf4c_v1, 2023.

[31] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 1251–1258, 2017. DOI: 10.1109/CVPR.2017.195.

[32] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, C. Wang, S. Ding, Y. Wu, L. Yuan. DF40: Toward next-generation deepfake detection. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Vancouver, Canada, Article number 925, 2024.

[33] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, 2019. DOI: 10.1109/ICCV.2019.00009.

[34] Z. Liu, P. Luo, X. Wang, X. Tang. Deep learning face attributes in the wild. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 3730–3738, 2015. DOI: 10.1109/ICCV.2015.425.

[35] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 3207–3216, 2020. DOI: 10.1109/CVPR42600.2020.00327.

[36] T. Karras, S. Laine, T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, USA, pp. 4396–4405, 2019. DOI: 10.1109/CVPR.2019.00453.

[37] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, S. Xie. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 16133–16142, 2023. DOI: 10.1109/CVPR52729.2023.01548.

[38] M. Tan, Q. V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, USA, pp. 6105–6114, 2019.

[39] M. Bora, T. Dhamija, S. Reddy, B. Chopin, P. Balaji, A. Das, A. Dantcheva. Authentica, IEEE Dataport, 2025.

DOI: 10.21227/nk22-0248.

[40] J. Liu, J. Wang, S. Hou, M. Ren, H. Wu, L. Ma, R. Pei, Z. He. Beyond face swapping: A diffusion-based digital human benchmark for multimodal deepfake detection, [Online], Available: https://arxiv.org/abs/2505.16512, 2025.

[41] J. P. Robinson, C. Qin, Y. Henon, S. Timoner, Y. Fu. Balancing biases and preserving privacy on balanced faces in the wild. *IEEE Transactions on Image Processing*, vol. 32, pp. 4365–4377, 2023. DOI: 10.1109/TIP.2023.3282837.

[42] S. Sharma, A. Sood, V. Kumar. Deepfake Synthetic-20K Dataset, IEEE Dataport, 2024. DOI: 10.21227/67x4-9g14.

[43] UTKFace, [Online], Available: https://susanqq.github.io/UTKFace/, 2017.

[44] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 5203–5212, 2020. DOI: 10.1109/CVPR42600.2020.00525.

[45] Z. Guo, Y. Liu, J. Zhang, H. Zheng, S. Shan. Face forgery detection with elaborate backbone, [Online], Available: https://arxiv.org/abs/2409.16945, 2024.

[46] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, USA, pp. 484–492, 2020. DOI: 10.1145/3394171.3413532.

[47] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, pp. 87–102, 2016. DOI: 10.1007/978-3-319-46487-9_6.

[48] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, E. Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 4873–4882, 2016. DOI: 10.1109/CVPR.2016.527.

[49] M. Jia, L. Tang, B. C. Chen, C. Cardie, S. Belongie, B. Hariharan, S. N. Lim. Visual prompt tuning. In *Proceedings of the 17th European Conference on Computer Vision*, Tel Aviv, Israel, pp. 709–727, 2022. DOI: 10.1007/978-3-031-19827-4_41.

[50] 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*, pp. 8748–8763, 2021.

[51] U. Ojha, Y. Li, Y. J. Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 24480–24489, 2023. DOI: 10.1109/CVPR52729.2023.02345.

[52] S. Gupta, A. Gupta. Model merging using geometric median of task vectors. In *Proceedings of the 38th Conference on Neural Information Processing Systems*, Vancouver, Canada, 2024.

[53] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, P. Bojanowski. DINOv3, [Online], Available: https://arxiv.org/abs/2508.10104, 2025.

[54] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 2818–2826, 2016. DOI: 10.1109/CVPR.2016.308.

[55] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, 2018. DOI: 10.1109/CVPR.2018.00474.

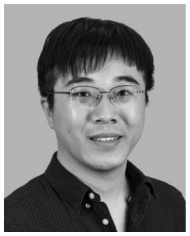
[56] M. Karki. Deepfake and real images, [Online], Available: https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images/data, 2023.

[57] Z. Yan, J. Wang, P. Jin, K. Y. Zhang, C. Liu, S. Chen, T. Yao, S. Ding, B. Wu, L. Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning*, Vancouver, Canada, 2025.

[58] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, N. Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, USA, pp. 772–781, 2021. DOI: 10.1109/CVPR46437.2021.00083.

[59] S. Ni, D. Meng, C. Yu, C. Quan, D. Ren, Y. Zhao. CORE: Consistent representation learning for face forgery detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 12–21, 2022. DOI: 10.1109/CVPRW56347.2022.00011.

[60] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the 16th European Conference on Computer Vision*, Glasgow, UK, pp. 86–103, 2020. DOI: 10.1007/978-3-030-58610-2_6.



Shu Hu received the M. Eng. degree in software engineering from the University of Science and Technology of China, China in 2016, the MA degree in mathematics from University at Albany, State University of New York (SUNY), USA in 2020, and the Ph. D. degree in computer science and engineering from University at Buffalo, SUNY, USA in 2022. He is an assistant professor and the Director of the Purdue Machine Learning and Media Forensics (M2) Lab in the School of Applied and Creative Computing, Purdue University, USA. He was a postdoc at Carnegie Mellon University, USA. He is a member of IEEE.

His research interests include machine learning, multimedia forensics, and computer vision.

E-mail: hu968@purdue.edu (Corresponding author)
ORCID iD: 0000-0003-1446-4140

Li Lin received the B. Sc. degree in communication engineering from Chongqing University, China in 2020. She is a Ph. D. degree candidate in the Department of Computer Science, Purdue University, USA.

Her research interests include computer vision, digital media forensics, and deep learning.

E-mail: lin1785@purdue.edu

Shail Desai received the B. Sc. degree in computer and information technology (CIT) from Purdue University, USA. He has over two years of experience in software engineering and data analysis, with hands-on work across multiple web development projects. He has a strong interest in research and believe that, in today′s AI-driven world, continuous research is essential to discovering new opportunities and driving technological innovation.

His research interests include multimedia forensics and computer vision.

E-mail: shdesai@purdue.edu

Aditya Pawar is an undergraduate student in computer science at Purdue University, USA. His work focuses on building practical, user-centric software solutions and exploring machine intelligence methods for real-world applications. He has experience working on research projects across academia and industry, including applied software engineering and data-driven systems.

His research interests include machine learning, web development, and AI systems.

E-mail: aspawar@purdue.edu

Guangyu Lin is a graduate student.

His research interests include machine learning and computer vision, particularly in AI-generated content detection and fairness-related issues.

E-mail: lin2285@purdue.edu

Xin Wang received the Ph. D. degree in computer science from the University at Albany, State University of New York (SUNY), USA in 2015. He is an assistant professor with the College of Integrated Health Sciences, and the AI Plus Institute at the University at Albany, SUNY, USA. He is a senior member of IEEE.

His research interests include machine learning, reinforcement learning, multimedia, and trustworthy AI.

E-mail: xwang56@albany.edu

Daniel S. Schiff received the A. B. degree in philosophy from Princeton University, USA in 2012, the M. Sc. degree in social policy from the University of Pennsylvania, USA in 2013, and the Ph. D. degree in public policy from the Georgia Institute of Technology, USA in 2022. He is currently an assistant professor in the Department of Political Science at Purdue University, USA, and serves as Co-Director of the Governance and Responsible AI Lab (GRAIL). His research examines the formal and informal governance of artificial intelligence through policy and industry, as well as AI′s social and ethical implications across domains such as education, labor, misinformation, and criminal justice. He is the founder and co-chair of the AI Community of the Association for Public Policy Analysis and Management (APPAM), serves on the Steering Committee of the Institute for Physical AI (IPAI), and serves on the Steering Committee of the Academic Alliance for AI Policy (AAAIP), a national forum connecting scholars with policymakers on AI governance. He is a member of the American Association for the Advancement of Science (AAAS), American Political Science Association (APSA), Association for the Advancement of Artificial Intelligence (AAAI), Association for Computing Machinery (ACM), Association for Public Policy Analysis and Management (APPAM), and Institute of Electrical and Electronics Engineers (IEEE).

E-mail: dschiff@purdue.edu

Sachi Nandan Mohanty received the Ph. D. degree from the Indian Institute of Technology Kharagpur, India in 2015 under the prestigious MHRD scholarship, followed by a Postdoctoral Fellowship at Indian Institute of Technology Kanpur, India in 2019. He has authored and edited fifty-two books, which are published by leading international publishers including IEEE-Wiley, Springer, CRC Press, NOVA, and De-Gruyter. He is a Fellow of the Institution of Engineers, IETE, Ambassador of the European Alliance for Innovation (Springer), Member in INAE, and Senior Member of IEEE Computer Society, Hyderabad Chapter.

His research interests include data mining, big data analysis, cognitive science, fuzzy decision making, brain-computer interfaces, cognition, and computational intelligence.

E-mail: sachinandan09@gmail.com

Ryan Ofman received the B.Sc. degree from Yale University, USA, where he focused on machine learning applications in astrophysics. He is the Head of AI Research at Deep Media AI. His research includes work on applying AI classification methods to galaxy morphology. At deep media AI, he leads research initiatives and collaborates with industry partners and academic institutions globally to advance the state of the

art in AI content authentication and ethical AI deployment.

His research interests include deepfake detection, multimedia forensics, and promoting equitable use of AI tools.

E-mail: ryan@deepmedia.ai

Narcis Bejtic is the Head of Operations at Originality.AI – the most accurate AI content detector and modern plagiarism checker on the market. After managing operations and business development for multiple successful businesses in the past, he joined Originality. AI and has helped the company be leader in AI detection.

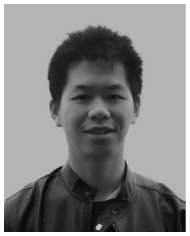
His research interest is AI-generated text detection.

E-mail: narcis@originality.ai

Jon Gillham is the Founder and CEO of Originality.AI. He has over a decade of experience in large-scale web publishing and content analysis, and was an early adopter of generative AI tools for content creation prior to the public release of ChatGPT. His work focuses on content originality, AI-generated media detection, and transparency in AI-assisted publishing systems.

His research interest is AI-generated text detection.

E-mail: jon@originality.ai



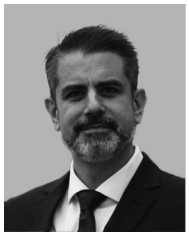
Wenbin Zhang is an assistant professor at Michigan Technological University, USA, and an Associate Member of the Te Ipu o Te Mahara Artificial Intelligence Institute, New Zealand. His research focuses on the theoretical foundations of machine learning, with an emphasis on responsible and socially beneficial AI. He has applied his work across multiple domains, including healthcare, digital forensics, energy, transportation, and finance. He actively contributes to the AI and interdisciplinary communities through leadership roles on organizing committees and editorial boards of leading venues, including as Volunteer Chair for WSDM′24 and Associate Editor for ACM Computing Surveys.

E-mail: wenbinzh@mtu.edu

Baoyuan Wu received the Ph. D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China in 2014. He is a tenured associate professor, and assistant dean (research) of School of Data Science, the Chinese University of Hong Kong, Shenzhen, China. He is a senior member of IEEE.

His research interests include trustworthy AI, generative AI, machine learning, computer vision, optimization, such as adversarial examples, backdoor learning, federated learning, face image editing/manipulation/generation, deepfake detection, etc.

E-mail: wubaoyuan@cuhk.edu.cn



Cristian Canton received the Ph. D. and M. Sc. degrees from Technical University of Catalonia, Barcelona, and completed his M. S. thesis at Swiss Federal Technology Institute of Lausanne (EPFL) on computer vision topics. He is director of the Barcelona Supercomputing Center, the largest research center in Spain. Formerly, head of Responsible AI at Meta, leading the company efforts around fairness, robustness and safety, governance and accountability, and transparency, specially in the field of Generative AI. Previously at Meta, he led the computer vision team within the Community Integrity division focused on harmful content, misinformation and manipulated media. From 2012–2016, he was at Microsoft Research in Redmond (USA) and Cambridge (UK); from 2009-2012, he was at Vicon (Oxford), bringing CV to produce visual effects for the cinema industry.

E-mail: cristian.canton@bsc.es

Xiaoming Liu received the Ph. D. degree in electrical and computer engineering from Carnegie Mellon University, USA in 2004. He is a distinguished professor at the Department of Computer Science of the University of North Carolina at Chapel Hill, USA. Before joining MSU in Fall 2012, he was a research scientist at General Electric (GE) Global Research. As a co-author, he is a recipient of Best Industry Related Paper Award runner-up at ICPR 2014, Best Student Paper Award at WACV 2012 and 2014, Best Poster Award at BMVC 2015, and Michigan State University College of Engineering Withrow Endowed Distinguished Scholar Award. He has been the Area Chair for numerous conferences, including CVPR, ICCV, ECCV, ICLR, NeurIPS, the Program CO-Chair of WACV′18, BTAS′18, IJCB′22, AVSS′22 conferences, and General Co-Chair of FG′23 conference. He is an Associate Editor of *Pattern Recognition* and *IEEE Transactions on Image Processing*. He has authored more than 150 scientific publications, and has filed 29 U.S. patents. He is a Fellow of IEEE and IAPR.

His research interests include computer vision, machine learning, and biometrics.

E-mail: liuxm@cs.unc.edu

Luisa Verdoliva is full professor with the Department of Electrical Engineering and Information Technology at University Federico II of Naples, Italy. She is the EiC of *IEEE Transactions on Information Forensics and Security* (2025–2027) and has been Chair of the IEEE IFS TC (2021–2022). She is the recipient of the IEEE SPS Best Paper Award (2025), Frontiers of Science Award (2025), the Google Research Award for Machine Perception (2018) and a TUM-IAS Hans Fischer Senior Fellowship (2020–2024). She is IEEE Fellow and 2026

NVIDIA Research Faculty Fellow.

Her research interests are in the field of image and video processing, with main contributions in the area of multimedia forensics.

E-mail: verdoliv@unina.it

Siwei Lyu is a SUNY Distinguished and Empire Innovation Professor of Computer Science and Engineering at the University at Buffalo, where he serves as Director of the Institute for Artificial Intelligence and Data Science and founding Director of the UB Media Forensic Lab. He has over 250 publications and support from NSF, DARPA, and DHS. He is a Fellow of the IEEE, IAPR, and AAIA, and a recipient of multiple research and innovation awards.

His research interests include media forensics, computer vision, and machine learning.

E-mail: siweilyu@buffalo.edu

Yongwei Tang is with Ant International, Shanghai, China. The research interests include machine learning and AI.

E-mail: tangyongwei.tyw@ant-intl.com

Zhiqiang Wu is with Ant International, Shanghai, China. The research interests include machine learning and AI.

E-mail: hanli.wzq@ant-intl.com

Jiawen Seow is with Ant International, Shanghai, China. The research interests include machine learning and AI.

E-mail: jiawen.seow@ant-intl.com

Zara Alaverdyan received the M. Sc. degree in data mining and knowledge management from Polytech Nantes, France, and the Ph. D. degree in machine learning for medical imaging from INSA Lyon, France in 2019. Currently, she works as applied scientist at Entrust, an identity verification company.

Her research interests include identity document understanding and fraud detection.

E-mail: Sarah.alaverdyan@entrust.com

Anne-Flore Baron is an applied scientist specializing in computer vision and machine learning, with an academic background from École polytechnique (Palaiseau, France) and the MVA Master′s program at ENS Paris-Saclay (Paris, France). She has led AI initiatives in agricultural imaging at Inarix and in identity verification at Entrust, focusing on the development of state-of-the-art, production-ready AI systems. Her prior research experience includes work on meta-learning for question answering at Mila (Montreal, Canada), domain adaptation and natural language processing at the TALia lab at onepoint (Paris, France), and Bayesian learning and graph-based credit scoring at Mentorica (Singapore).

E-mail: anneflore.contact@gmail.com

Simon Bozonnet received the M. Eng. degree from INSA Lyon, France in 2008 and the Ph. D. degree from EURECOM/TelecomParisTech, France in 2012. During his doctoral studies at EURECOM, his research focused on speech processing and media indexing. Key highlights of his work include the LIA-EURECOM submission to the NIST Rich Transcription 2009 evaluation and the "ACAV" project, which aimed to improve web accessibility for the visually and hearing impaired. In 2021, he joined Onfido (now part of Entrust), where he currently works as Applied Scientist.

His research interests include computer vision and document AI for information extraction and fraud detection used to build secure, automated identity-verification systems.

E-mail: Simon.bozonnet@entrust.com

Martins Bruveris is with Entrust Corp., Shakopee, USA. The research interests include machine learning and AI.

E-mail: martins.bruveris@entrust.com

Jochem Gietema is with Entrust Corp., Shakopee, USA. The research interests include machine learning and AI.

E-mail: jochem.gietema@entrust.com

Lucia Innocenti is with Entrust Corp., Shakopee, USA. The research interests include machine learning and AI.

E-mail: lucia.innocenti@entrust.com

Lisa Ivanova is with Entrust Corp., Shakopee, USA. The research interests include machine learning and AI.

E-mail: lisa.ivanova@entrust.com

Olivier Koch received the M. Eng. degree from ENSTA Paris, France and the Ph. D. degree from the Massachusetts Institute of Technology, USA. He leads the machine learning team at Onfido, acquired by Entrust. Prior to that, he was leading machine learning teams at Thales and Criteo in the fields of computer vision and recommender systems. He has published work at CVPR, ICCV, ICRA and IJFR. He teaches deep learning at ENSAE Paris, France.

E-mail: olivier.koch@entrust.com

Harry Ni is with Entrust Corp., Shakopee, USA. The research interests include machine learning and AI.

E-mail: harry.ni@entrust.com

Arthur Pajot is with Entrust Corp., Shakopee, USA. The research interests include machine learning and AI.

E-mail: arthur.pajot@entrust.com

Romain Sabathe is with Entrust Corp., Shakopee, USA. The research interests include machine learning and AI.

E-mail: romain.sabathe@entrust.com

Fengming Gu received the B. Sc. degree in computer science and technology from University of Chinese Academy of Sciences, China in 2022. He is currently a Ph. D. degree candidate with the Institute of Computing Technology (ICT), Chinese Academy of Sciences, China.

His research interest is computer vision, particularly include deepfake detection and progressive defense.

E-mail: gufengming18@mails.ucas.ac.cn

Xingming Long received the B. Sc. degree in computer science from Tsinghua University, China in 2021. He is currently a Ph. D. degree candidate with the Institute of Computing Technology (ICT), Chinese Academy of Sciences, China.

His research interests include face anti-spoofing and domain generalization.

E-mail: xingming.long@vipl.ict.ac.cn

Jie Zhang received the Ph. D. degree from the University of Chinese Academy of Sciences (CAS), China. He is currently an associate professor with the Institute of Computing Technology, CAS, China.

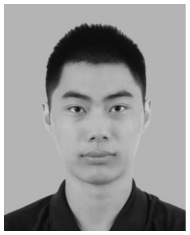
His research interests include computer vision, pattern recognition, machine learning, particularly include adversarial attacks and defenses, domain generalization, AI safety and trustworthiness.

E-mail: zhangjie@ict.ac.cn

Wenqing Ge received the B. Eng. degree in internet of things engineering from Qingdao University of Science and Technology, China in 2025. She is currently a master student at the College of Electronic Engineering, Ocean University of China, China.

Her research interests include face forgery detection and the fairness of face forgery detection.

E-mail: gwq17866836583@163.com



Xiangkui Cao received the B. Sc. degree from University of Chinese Academy of Sciences, China in 2023. He is currently a Ph. D. degree candidate from University of Chinese Academy of Sciences, China.

His research interest includes AI safety and truthfulness.

E-mail: xiangkui.cao@vipl.ict.ac.cn

Yuecong Min received the Ph. D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, China. He is currently a postdoctoral researcher with the Institute of Computing Technology, Chinese Academy of Sciences, China.

His research interests include computer vision, pattern recognition, and machine learning. He especially focuses on sign language processing, vision-language modeling and the related research topics.

E-mail: minyuecong@ict.ac.cn

Yingjie Liu received the B. Eng. degree in information engineering from Qingdao University of Science and Technology, China in 2022, and the M. Eng. degree in electronic information from Ocean University of China, China in 2025. He is currently a Ph. D. degree candidate at the College of Electronic Engineering, Ocean University of China, China.

His research interests include face forgery detection and presentation attack detection.

E-mail: yingjieliu@stu.ouc.edu.cn

Zonghui Guo received the Ph. D. degree in intelligence information and communication systems from the Ocean University of China, China in 2021. From 2021 to 2024, he was a postdoctoral researcher at the Institute of Computing Technology, Chinese Academy of Sciences (CAS), China. In 2025, he joined the Ocean University of China, where he is currently an associate professor.

His research interests include face forgery detection, low-level vision applications such as image/video harmonization and image/video enhancement.

E-mail: guozonghui@ouc.edu.cn

Shiguang Shan received the Ph. D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), China in 2004. He has been a full professor with ICT since 2010, where he is currently the director of the Key Laboratory of Intelligent Information Processing, CAS, China. He has published more than 300 articles in related areas. He served as the General Co-Chair for IEEE Face and Gesture Recognition 2023, the General Co-Chair for Asian Conference on Computer Vision (ACCV) 2022, and the Area Chair of many international conferences, including CVPR, ICCV, AAAI, IJCAI, ACCV, ICPR, and FG. He was/is Associate Editors of several journals, including *IEEE Transactions on Image Processing*, *Neurocomputing*, *Computer Vision and Image Understanding*, and *Physical Review Letters*. He was a recipient of the China′s State Natural Science Award in 2015 and the China′s State S&T Progress Award in 2005 for his research work. He is a fellow of IEEE.

His research interests include signal processing, computer vision, pattern recognition, and machine learning.

E-mail: sgshan@ict.ac.cn

Jinhee Park received the B. Sc. and M. Sc. degrees in computer science from Chung-Ang University (CAU), Seoul, Republic of Korea in 2019 and 2021, respectively, where she is currently a Ph. D. degree candidate in computer science.

Her research interests include metric learning and confidence calibration.

E-mail: iv4084em@keti.re.kr

Minjun Kim received the B. Sc. degree from the Department of AI at Chung-Ang University, Republic of Korea. He is currently a master student in the Computer Vision Laboratory (CVLab) at POSTECH, under the supervision of Professor Suha Kwak.

His research interest is computer vision.

E-mail: ekdh635@cau.ac.kr

Ahyeon Park studied applied statistics with a double major in artificial intelligence at Chung-Ang University, Republic of Korea, and is currently a Clinical Research Coordinator in the Division of Cardiology at Chung-Ang University Hospital, Republic of Korea.

E-mail: 423data@gmail.com

Guisik Kim received the M. Sc. degree in low level vision (supervised by Prof. Junseok Kwon) and the Ph. D. degree in computer science and engineering from Chung-Ang University (CAU), Republic of Korea in 2017 and 2023, respectively. He is currently a full-time researcher with Korea Electronics Technology Institute (KETI).

His research interests include low-level vision, satellite image processing, and deepfake detection.

E-mail: specialre@naver.com

Taewoo Kim received the B. Sc. degree from the Department of Electronics and Communications Engineering, Kwangwoon University, Republic of Korea in 2018, and the M. Sc. degree from the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Republic of Korea in 2020. From 2020 to 2023, he was a researcher with NCSOFT AI Center, Republic of Korea. Since 2023, he has been a senior researcher with the Multi-Modal Research Center, Korea Electronics Technology Institute (KETI).

His research interests include speech and audio signal generation and audio deepfake detection.

E-mail: kimtaewoo@keti.re.kr

YoungJoon Yoo received the B. Sc. degree in electrical and computer engineering at Seoul National University, Republic of Korea in 2011, and the Ph. D. degree in electrical engineering from Seoul National University, Republic of Korea in 2017. He was a research scientist in NAVER AI Research, and also leaded the Image Vision team, NAVER CLOVA. Currently, he is an assistant professor of the Department of Artificial Intelligence, Chung-Ang University, Republic of Korea.

His research interests include deep learning for computer vision and probabilistic theory.

E-mail: yjyoo3312@cau.ac.kr

Junseok Kwon received the B. Sc. degree in 2006 from the Seoul National University, Republic of Korea in 2006, the M. Sc. degree in 2008 in the topic of object tracking (supervised by Prof. Kyoung Mu Lee), and the Ph. D. degree in electrical engineering and computer science in 2013. He is a professor in the School of Computer Science and Engineering at Chung-Ang University, Republic of Korea. He is working in the field of object tracking to capture the dynamics of cities.

His research interests include visual tracking, visual surveillance, and Monte Carlo sampling method.

E-mail: jskwon@cau.ac.kr

Zhaoda Li primarily works on algorithms related to the trustworthiness and risk control of AI systems, including data quality, generated content detection, and the identification and defense against model misuse.

E-mail: zdaiot@163.com

Mengyun Tang is a freelance researcher specializing in AI security, including large language model security, model misuse, data leakage, and AI-generated content detection. She is a freelance researcher specializing in AI security, including large language model security, model misuse, data leakage, and AI-generated content detection.

E-mail: merrytangmengyun@gmail.com

Leyang Huang is a CS undergraduate at The University of Hong Kong, China, with a deep passion for generative AI and AI security. Her journey so far has been a thrilling ride, collaborating with both academic and industry experts to tackle real-world challenges using AI, and she is eager to contribute to its evolution in meaningful ways.

E-mail: hleyang@connect.hku.hk

Bogdan Dura received the B. Eng. degree in computer science engineering from the Faculty of Mathematics and Computer Science, University of Bucharest, Romania. He is currently a second-year master student in artificial intelligence at the University of Bucharest, Romania, and a software engineer at the National Institute of Research and Development in Informatics (ICI Bucharest).

His research interests include computer vision, deep learning and robotics.

E-mail: bogdan.dura@ici.ro

Sebastian Balmuş received the B. Eng. degree in telecommunications engineering from the Technical University of Cluj-Napoca, Romania in 2024. He is currently a second-year master student in Natural Language Processing at the University of Bucharest, Romania, and a software engineer at the National Institute of Research and Development in Informatics (ICI Bucharest).

His research interests include deep learning, computer vision and natural language processing.

E-mail: sebastian.balmus@ici.ro

Fang-Yi Su is a visiting scholar with the Department of Biomedical Informatics, Harvard Medical School, USA.

His research interests include AI for medicine, fairness in AI, and generative models.

E-mail: fang-yi_su@hms.harvard.edu

Tsung-Hua Lee is currently a Research Associate with the Department of Biomedical Informatics, Harvard Medical School, USA.

His research interests include agentic and multimodal AI.

E-mail: tsunghua_lee@hms.harvard.edu

Ting-Wan Kao received the M. D. and M. Sc. degrees in clinical medicine, with a background in translational cancer research and drug development targeting cancer recurrence, stemness, and therapy resistance. She is a graduate researcher in the Department of Biomedical Informatics at Harvard Medical School, USA.

Her research interests include computational pathology and machine learning for cancer risk prediction, with an emphasis on metastasis, recurrence, survival modeling, and algorithmic fairness across populations.

E-mail: tingwan_kao@hms.harvard.edu