# The Competition of Fairness in AI Face Detection

**Shu Hu**[*†] Xin Wang[†] Daniel S. Schiff[†] Sachi Nandan Mohanty [†] Ryan Ofman[†]
Narcis Bejtic[†] Jon Gillham[†] Wenbin Zhang[†] Baoyuan Wu[†] Cristian Canton[†]
Xiaoming Liu[†] Luisa Verdoliva[†] Siwei Lyu[†]
hu968@purdue.edu m2lab.purdue@gmail.com

## Abstract

This competition focuses on advancing fairness-aware detection of AI-generated (deepfake) faces and promoting new methodological innovations, addressing a critical gap where fairness methods developed in machine learning have been largely overlooked in deepfake detection. In the competition, participants will work with two large-scale datasets provided by the organizers: AI-Face (CVPR 2025), a million-scale, demographically annotated dataset for training and validation, and PDID (AAAI 2024), a newly curated dataset comprising real-world deepfake incidents, reserved for testing. Participants are tasked with developing models that achieve strong utility performance (e.g., AUC) while ensuring fairness generalization under real-world deployment conditions. The baseline method, PG-FDD (published at CVPR 2024 from the organizer's group), which demonstrates state-of-the-art performance in fairness generalization for AI face detection, will be provided to support participation. The competition's potential impact includes fostering the development of robust, fair, and generalizable deepfake detectors, raising awareness of fairness challenges in combating AI-generated fakes, and promoting responsible AI and machine learning deployment in societal applications such as media forensics and digital identity verification. Our competition is fortunately sponsored by Deep Media AI and Originality.AI companies. The challenge link is https://sites.google.com/view/aifacedetection/home.

**Keywords** AI-generated Face, Deepfake Detection, Fairness

## 1 Competition description

### 1.1 Background and impact

DeepFake technology, powered by advanced deep neural network (DNN)-based generative AI methods such as variational autoencoders [1], generative adversarial networks (GAN) [2], and diffusion models (DM) [3], produce 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 DeepFakes is especially alarming in sensitive areas, such as the 2024 US Presidential Election [4, 5, 6] or when they involve public figures such as Taylor Swift [7, 8]. Fig.1 Top shows several well-known real-world DeepFakes. Their 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.

[*]The Leader organizer should be the first author of the proposal.

[†]Leader and Backup(s) organizers should read and acknowledge Competition Chairs' communications.

Preprint. Under review.

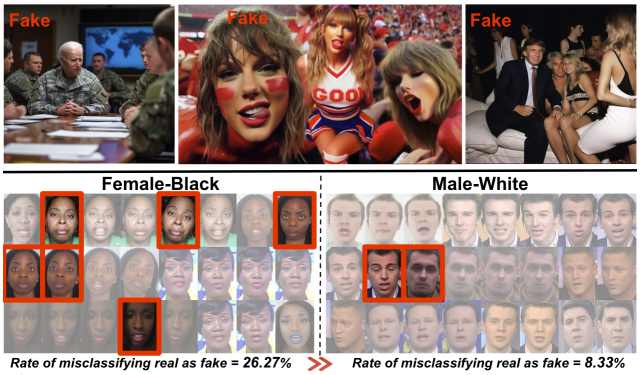

Figure 1: *(Top) Examples of several visual DeepFakes. (Bottom) The existence of unfairness in DeepFake detection. The red boxes highlight the wrong predictions.*

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, 10, 11, 12, 13, 14] have highlighted biases in DeepFake detection, such as higher accuracy for lighter-skinned deepfakes (see Fig.1 Bottom), but available solutions have been less forthcoming about the nature of these biases[15, 16].

While the machine learning community has proposed numerous fairness algorithms to help alleviate these biases [17], their application in deepfake detection remains limited. Moreover, naïvely 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.

Therefore, we propose to organize **the first competition focused on fairness in AI-generated face detection**, 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 more fair deepfake 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.

**Competition Objective:** 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.

**Relevance and Participants**. Our proposed competition is closely aligned with several core research topics highlighted in the NeurIPS 2025 Call for Papers, including *Social and Economic Aspects of Machine Learning, General Machine Learning, Evaluation, Deep Learning, and Applications*. By targeting fairness in AI-generated face detection, the competition directly engages with pressing questions at the intersection of algorithmic bias, robustness, and the deployment of deep learning systems in high-stakes environments. The relevance of these themes, combined with the widespread interest in generative AI and media forensics, positions the competition to attract significant attention from the NeurIPS community. We anticipate participation from over 200 researchers, practitioners, and students, and expect active engagement with both the competition and the affiliated workshop. The competition is designed to appeal to a broad audience, including those working in generative AI, robustness, fair learning, and AI security.

### 1.2 Novelty

To the best of our knowledge, this is the **first** competition specifically focused on fairness in the detection of AI-generated faces.

### 1.3 Data

The competition will rely on two proprietary datasets (AI-Face [18] and PDID [19]) provided by the organizers, comprising approximately **two million** face images in total. The organizers retain full ownership of both datasets and will make them freely available for use during the competition. The PDID dataset includes ground truth annotations that have not been previously released and have been maintained in a secure and confidential manner. A thorough review has been conducted, and no potential feature leakage issues have been identified in either dataset.

Detailed information about our datasets is provided below:

1. The **AI-Face** dataset [18], developed by organizer Dr. Shu Hu's research group, is a million-scale, demographically annotated dataset that will be provided to participants as the primary training and validation sets for model development. The associated research paper has been accepted for publication at CVPR 2025. The dataset and related resources can be accessed via the following GitHub repository: https://github.com/Purdue-M2/AI-Face-FairnessBench.
   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 the authors. Further details are available in [18].
   Most importantly, **the dataset collection and annotation generation are approved by Purdue's Institutional Review Board.** The dataset is only for research purposes. All data are sourced from publicly available datasets, and we strictly comply with each dataset's license agreement to ensure lawful inclusion and permissible secondary use for training. Our annotation processes prioritize ethical considerations: 1) 76% images we annotated are generated facial images, ensuring no potential for harm to any individual. 2) For real images, we only provide annotations for content either licensed by the original copyright holders or explicitly stated as freely shareable for research purposes.
2. The **PDID** dataset [19], provided by co-organizer Dr. Daniel S. Schiff, will serve as the test set for the competition. The associated research paper was accepted at AAAI 2024. The dataset is hosted on the AirTable platform and can be accessed at http://bit.ly/pdid.
   Data collection for the PDID dataset began in June 2023 and remains ongoing. The dataset currently includes deepfake incidents dating back to 2017. To focus on cases that have gained public attention, data are primarily sourced from English-language social media posts and widely circulated news websites. The collection process starts with media reports on prominent deepfake incidents and employs a snowball sampling strategy to gather related articles and social media content. After rigorous data cleaning, the dataset now comprises approximately 9,500 real images and 208,900 AI-generated images. Demographic annotations are inferred using the same automated annotators developed for the AI-Face dataset. As this dataset will be reserved exclusively for evaluation purposes, the demographic labels will remain confidential and will be used solely to assess the performance and fairness of participants' submitted models in the organizers' side.

We confirm that the data complies with all data-related concerns of the NeurIPS Code of Ethics, including privacy, consent, deprecation, copyright, and fair use.

### 1.4 Tasks and application scenarios

In our competition, we plan to include a singular task, reflecting a significant real-world application scenario, as detailed below.

**Task** : Improving Fairness Generalization in AI Face Detection.

*Real-world Scenario*: In our competition, the AI-Face training set will represent 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 deepfakes, such as those collected in the PDID database. Real-world deepfakes 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.

*Justify the Problem*: 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. We are currently in an era of rapid advancements in generative AI, with the continual release of powerful models such as GPT-4o. These models can be readily misused by deepfake makers to create realistic AI-generated faces, exacerbating societal risks. The dynamic nature of generative technologies creates a persistent cat-and-mouse game between deepfake generation and detection, highlighting the urgent need for fairness solutions that generalize effectively to novel threats.

Therefore, this task aims to encourage participants to develop more advanced, generalizable fair deepfake detectors, with the goal of creating potential solutions to mitigate these emerging challenges.

We confirm that both tasks and application scenarios comply with the NeurIPS Code of Ethics.

### 1.5 Metrics

Following our recent fairness benchmark work [18] for detecting AI-generated faces, we will consider 4 fairness metrics commonly used in the fairness community [20, 17, 21, 22, 23] and 5 more widely used utility metrics [24, 25, 26, 27]. For *fairness* metrics, we consider Demographic Parity ($F_{DP}$) [20, 17], Max Equalized Odds ($F_{MEO}$) [22], Equal Odds ($F_{EO}$) [21], and Overall Accuracy Equality ($F_{OAE}$) [22] 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 (Dark), Male-Light (M-L), Male-Medium (M-M), and Male-Dark (M-D), where we group 10 categories of skin tones into Light (Tone 1-3), Medium (Tone 4-6), and Dark (Tone 7-10) for simplicity according to [28]. 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).

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:

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

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 will only consider detectors that demonstrate superior utility performance (e.g., AUC) compared to the provided baselines. If multiple participants achieve the same average fairness ranks (Avg-$F_R$), we will compare their primary utility (e.g., 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 (e.g., EER) and recalculate their Avg-$F_R$ for comparison. This process will be repeated iteratively until the best detector is identified.

We anticipate that this procedure will allow us to determine a winner without exhausting all available utility metrics. However, in the unlikely event that multiple detectors achieve identical best Avg-$F_R$ across all utility metrics, we have a preset contingency plan. In such a case, each team will be required to submit a one-page summary describing their developed detector. A final decision will then be made through a majority vote by an odd-numbered subset of randomly selected members from the organizing committee. We emphasize that the likelihood of this worst-case scenario occurring is extremely low.

### 1.6 Baselines, code, and material provided

**Baseline**. As demonstrated in our benchmark study [18], PG-FDD [16] achieves the best overall fairness and utility performance. PG-FDD is also the first method specifically designed to improve fairness generalization in AI-generated face detection from the organizer Prof. Shu Hu's group. Accordingly, we will provide this model as the baseline for the competition. Preliminary results on the AI-Face dataset, including the performance of PG-FDD, are presented in Table 4 of [18].

**Code & Data**. The baseline code is publicly available at https://github.com/Purdue-M2/Fairness-Generalization and is also included in https://github.com/Purdue-M2/

AI-Face-FairnessBench. The data-loading tools necessary for handling the datasets are provided within the same repository. Both the baseline code and data-loading tools will be accessible prior to the start of the competition.

For the AI-Face dataset, the face images can be downloaded from https://purdue0-my.sharepoint.com/personal/lin1785_purdue_edu/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Flin1785%5Fpurdue%5Fedu%2FDocuments%2FAI%5FFace%5FimagesV2&ga=1. To access the corresponding annotations, 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. Detailed instructions for this process are outlined at https://github.com/Purdue-M2/AI-Face-FairnessBench?tab=readme-ov-file#download.

### 1.7 Website, tutorial and documentation

The competition website is available at https://sites.google.com/view/aifacedetection/home. On the website, participants can find a dedicated FAQ page, which will be regularly updated as new questions are received to ensure clear communication. Our official email address (m2lab.purdue@gmail.com) is also provided on the website and can be used by participants to reach the organizing team with any inquiries. The website is currently online and will be refined with additional details and updates as the competition progresses. Participants can learn more about the targeted problem addressed in this competition by referring to the following reference:

[CVPR'25] Li Lin, Santosh, Mingyang Wu, Xin Wang, Shu Hu. AI-Face: A Million-Scale Demographically Annotated AI-Generated Face Dataset and Fairness Benchmark. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 2025, 6.

[CVPR'24] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, Shu Hu. Preserving Fairness Generalization in Deepfake Detection. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2024, 6.

## 2 Organizational aspects

### 2.1 Protocol

Participants will join the competition by registering through the Codabench platform (https://www.codabench.org/competitions/7166/?secret_key=bf3e8845-dadc-4472-98aa-5b4f022cb308) using their real identity and affiliation. Once registered, they can access the AI-Face dataset for training and validation via a provided download link. To obtain the annotated demographic labels, participants must agree to the End-User License Agreement (EULA). Submissions will be in the form of result files (not code), specifically score files with each line representing the image ID and a confidence score indicating how likely the image is AI-generated. All submissions will be uploaded directly to Codabench, where an automated evaluation script will compute both utility performance and fairness performance.

The competition consists of two phases: a validation phase and a testing phase. During the validation phase, participants receive feedback on a subset of the test set. In the testing phase, participants submit predictions for a completely held-out test set, with the final leaderboard determined by private evaluation. To prevent overfitting and cheating, the test set labels are never released, submission limits are enforced, and top teams will be required to open-source their code and submit a detailed methodology report. The Codabench platform ensures the secure handling of submissions, real-time leaderboard updates, and fair evaluation. Prior to the competition launch, the organizers will conduct a full internal beta test to validate the submission pipeline, scoring program, and leaderboard logic.

### 2.2 Rules and Engagement

**Rules.**

1. All participants should register for this challenge with their real names, affiliations (including department, full name of university/institute/company, country), and affiliation E-mails. Incom-

plete and redundant registrations will be removed without notice. Each team can have at most ten people.

2. All participants should submit a complete solution to this challenge during the validation and testing phase. A complete solution includes a score file and a qualified methodology paper (only for the testing phase).
   The format of a score file is as follows:

   $$\textit{imageID <TAB> score},$$

   where *imageID* is the id of the test file, and score is a numerical value – a higher value for fake images and lower value for fake images. A sample of the file is shown here:

   $$\begin{array}{ll} \textit{000001} & \textit{0.9915} \\ \textit{000002} & \textit{0.2418} \\ \ldots & \\ \textit{100003} & \textit{0.8718} \end{array}$$

   Each participating team should submit a qualified methodology paper that includes:
   - How the data is used.
   - Which model(s) are used and the corresponding model parameters.
   - How the models are trained (i.e., loss function, optimizer, learning rate, batch size, stopping criteria, etc.)
   - The description of the pre-trained models (if the participants have used any publicly available pre-trained models for embedding extraction).
   - Number of model parameters (trainable and non-trainable).
   - Time (secs) required to process one image.
   - Performance on the validation set.
3. All participants should agree that the submitted papers can be publicly available to the community on the challenge website and proceeding, and organizers can use the information provided by the participants, including scores and papers.
4. All participants should agree to make their code and models publicly available to the community for reproduction if they are in the top 10 teams on the final testing leaderboard.
5. Participants are not allowed to register multiple teams and accounts. Participants from the same research group are also not allowed to register multiple teams. The organizers keep the right to disqualify such participants.
6. Redistribution or transfer of data or data link is not allowed. Participants should use the data only by themselves.
7. Pretrained models and publicly available datasets are allowed to use.

**Discussion of the rules**. The competition rules are carefully designed to promote fairness, transparency, and scientific rigor, which are essential for achieving the goals of this challenge. By requiring participants to register with verified personal and institutional information, the competition prevents duplicate or anonymous entries that could compromise the integrity of the evaluation process. The requirement for complete solution submissions—including both score files and detailed methodology papers—enables rigorous assessment of each team's approach, which fosters reproducibility and enables meaningful comparisons across different methods. Standardizing the score file format and requesting comprehensive documentation of model development and training procedures ensures that participants provide sufficient technical details, which supports robust evaluation and knowledge dissemination. Additionally, the promotion of the public release of papers and, for top-performing teams, code and models aligns with the broader goal of advancing community-driven research and ensuring that the outcomes of the competition are reproducible and extendable. Prohibiting multiple registrations from the same individuals or research groups further promotes fairness and prevents manipulation of rankings. Restrictions on the redistribution of data protect the confidentiality and integrity of our datasets, while the allowance for using publicly available pretrained models and datasets encourages innovation without unduly restricting participants' methodological choices.

**Communication**. We have provided our official email address (`m2lab.purdue@gmail.com`) on the challenge website, which participants can use for all official communications with us. This email address will also be used to send reminder messages to all registered participants. If any rules or deadlines are modified, we will first post the updates on the challenge website. Simultaneously, we will send a notification email to all participants to ensure that they are promptly informed of the changes.

### 2.3 Schedule and readiness

(Ready) 2025-04-13 The organizing committee has been finalized, the competition website has been launched, and the CodaLab platform has been initialized.

(Preparing) 2025-05-15 Registration open. We will distribute the information of our competition. The plan is detailed in Section 2.4.

(Preparing) 2025-07-01 Registration close.

(Preparing) 2025-07-2 Release of starting kit (training and validation sets are included) and Validation submission on CodaBench will open.

(Preparing) 2025-10-2 Validation submission on CodaBench will close

(Preparing) 2025-10-3 Test set is released and testing submission on CodaBench will open

(Preparing) 2025-10-31 End of testing submission

(Preparing) 2025-11-10 Top teams are invited to submit papers and top three teams will be invited to give oral presentations.

### 2.4 Competition promotion and incentives

Following the strategies from our 1st AIMS workshop in 2024 [29], we design a structured plan to announce or invite participants to our workshop:

1. **Through Academic and Professional Networks**: We will prepare a flyer about the competition, which includes the topic, guidelines for participants, and important dates. The organization committee members will distribute the flyer in their field. We will contact professional associations like IEEE and ACM to list our competition in their event calendars or newsletters. Note that our organization committee members include three **IEEE Fellows** and **ACM Distinguish Members**. We will also use LinkedIN groups, X, and Facebook to promote our competition. Create a hashtag specific to our competition for easy tracking and visibility.
2. **Through the Collaboration with Institutions and Industry**. We will reach out to universities and research institutions with strong programs in computer science, social science, and AI, inviting faculty and students to participate. We will contact companies and startups in the tech sector, especially those focusing on AI, computer vision, and multimedia security, to sponsor or participate in our competition. **For instance, a well known Deepfake Detection company, Deep Media AI, has sponsored this competition, providing winners prizes of up to $900 cash and 1000 dollars in credits to a commercial deepfake detection platform. Originality.AI company also sponsored this competition, providing winners $900 credits to use their detection platform.**
3. **Through Specialized Forums**: We will post details about our competition on specialized forums such as Reddit (e.g., r/MachineLearning, r/computervision) [30], Stack Exchange communities [31], and specific research groups on ResearchGate [32] or Machine Learning News Google Groups [33].
4. **Direct Invitations**: We plan to personally invite respected researchers in the field to give talks. Their commitment can attract more attendees. In addition, We will reach out to our 1st AIMS workshop participants directly through email invitations, encouraging them to join the competition and share the information within their networks.

At the current stage, we plan to invite participants to submit a paper describing their methods if their model performance exceeds that of the baseline models. Selected papers will be given the opportunity to be published in our challenge proceeding (in preparing). As reference, examples of our previous book publication records can be found at https://link.springer.com/book/9783031917974?utm_medium=catalog&utm_source=sn-bks&utm_campaign=search_tool&utm_content=my_flyer#overview and https://link.springer.com/book/9783031907227?srsltid=AfmBOorCfNVrqjiJXqzExg5WV5KDyVZM4KAkaNuBJPRKgEBbVdeihyJX. There will be no specific requirements regarding the level of authorship for the submitted papers.

Most importantly, we have two competition sponsors:

1. Deep Media AI (https://deepmedia.ai/), provides winners prizes: 1st Place: $500 Cash +Deepfake Detection credits (worth $1000) to DeepID (i.e., their product); 2nd Place: $300 Cash +

Deepfake Detection credits (worth $500) to DeepID; 3rd Place: $100 Cash + Deepfake Detection credits (worth $200) to DeepID. All participants will obtain 10 free scans to DeepID.
2. Originality.AI (https://originality.ai/) provides winners prizes: 1st Place: credits (worth $500) to originality.ai (i.e., their product); 2nd Place: credits (worth $300) to originality.ai; 3rd Place: credits (worth $100) to originality.ai.

To attract participants of groups under-represented at NeurIPS, we will invite speakers and attendees from underrepresented groups. This includes reaching out to organizations, universities, and communities known for supporting diversity in science and engineering. We will partner with organizations and societies focused on diversity in STEM (e.g., National Society of Black Engineers [34], Society of Women Engineers [35], Association for Women in Science [36]) to promote the workshop and identify potential speakers and attendees.

# 3 Resources

## 3.1 Organizing team

The organizing team is comprised of experts with extensive and complementary experience in their respective fields such as Machine Learning (Shu Hu, Wenbin Zhang), Fairness (Shu Hu, Daniel S. Schiff, Wenbin Zhang, Cristian Canton, Xiaoming Liu), AI (Xin Wang, Wenbin Zhang, Ryan Ofman, Narcis Bejtic, Jon Gillham), Political Science (Daniel S. Schiff), Data Mining (Sachi Nandan Mohanty), Media Forensics (Shu Hu, Luisa Verdoliva, Siwei Lyu, Ryan Ofman, Narcis Bejtic, Jon Gillham), and Computer Vision (Cristian Canton, Xiaoming Liu). The team member roles are demonstrated below:

1. **Coordinators**: Prof. Shu Hu
2. **Data Providers**: Prof. Shu Hu, Prof. Daniel S. Schiff
3. **Platform Administrators**: Prof. Xin Wang, Prof. Wenbin Zhang
4. **Baseline Method Providers**: Prof. Shu Hu
5. **Publication Chair**: Prof. Sachi Nandan Mohanty
6. **Beta Testers**: Prof. Xin Wang, Prof. Wenbin Zhang, Prof. Baoyuan Wu
7. **Evaluators**: Prof. Baoyuan Wu, Dr. Cristian Canton, Prof. Xiaoming Liu, Prof. Luisa Verdoliva, Prof. Siwei Lyu
8. **Sponsor**: Mr. Ryan Ofman (Deep Media AI), Mr. Narcis Bejtic (Originality.AI), Mr. Jon Gillham (Originality.AI)

*To enhance team collaboration, we plan bi-weekly online meetings to monitor competition preparation progress and discuss the next step plans.*

## 3.2 Resources provided by organizers

N/A.

## 3.3 Support requested

Given that the NeurIPS 2025 Competition Track is an in-person event, we respectfully request the following support from the conference:

1. We request **a waiver of the registration fees** for organizers and a few student volunteers from the organizers' laboratories, to support the effective management and smooth operation of the competition.
2. We request the allocation of a **dedicated physical space** (with a tolerance of 200 audiences) at the conference venue to host our competition.
3. Audio-visual equipment, including **projectors, microphones, and poster boards**, would be essential for facilitating presentations and interactive sessions.
4. We request an additional table in the competition room for our sponsors (i.e., Deep Media AI, Originality.AI) to show product demos.

## References


[1] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. Advances in neural information processing systems, 33:19667–19679, 2020.

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8110–8119, 2020.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34:8780–8794, 2021.

[4] Image of biden planning military action in fatigues is fake. In https://factcheck.afp.com/doc.afp.com.34H74GF.

[5] Trump's dad resurrected via ai to tell son he's a disgrace. In https://futurism.com/the-byte/trump-dad-ai.

[6] Trump says red marks on hands may have been ai. In https://thehill.com/homenews/campaign/4441928-trump-says-red-marks-on-hands-may-have-been-ai/.

[7] Taylor swift, x and ai-generated images. In https://shorturl.at/0C2jB.

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

[9] Loc Trinh and Yan Liu. An examination of fairness of ai models for deepfake detection. IJCAI, 2021.

[10] Ying Xu, Philipp Terhörst, Kiran Raja, and Marius Pedersen. A comprehensive analysis of ai biases in deepfake detection with massively annotated databases. arXiv preprint arXiv:2208.05845, 2022.

[11] The 19th international conference on advanced video and signal-based surveillance. In https://www.avss2023.org.

[12] Aakash Varma Nadimpalli and Ajita Rattani. Gbdf: gender balanced deepfake dataset towards fair deepfake detection. arXiv preprint arXiv:2207.10246, 2022.

[13] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence, pages 1–53, 2022.

[14] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied intelligence, 53(4):3974–4026, 2023.

[15] Shu Hu, Yan Ju, Shan Jia, George H Chen, and Siwei Lyu. Improving fairness in deepfake detection. IEEE/CVF Winter Conference on Applications of Computer Vision, 2024.

[16] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[17] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6):1–35, 2021.

[18] Li Lin, Santosh, Mingyang Wu, Xin Wang, and Shu Hu. Ai-face: A million-scale demographically annotated ai-generated face dataset and fairness benchmark. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025.

[19] Christina P Walker, Daniel S Schiff, and Kaylyn Jackson Schiff. Merging ai incidents research with political misinformation research: Introducing the political deepfakes incidents database. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 23053–23058, 2024.

[20] Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. Ffb: A fair fairness benchmark for in-processing group fairness methods. In ICLR, 2024.

[21] Jialu Wang, Xin Eric Wang, and Yang Liu. Understanding instance-level impact of fairness constraints. In International Conference on Machine Learning, pages 23114–23130. PMLR, 2022.

[22] Hao Wang, Luxi He, Rui Gao, and Flavio P Calmon. Aleatoric and epistemic discrimination in classification. ICML, 2023.

[23] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference, pages 214–226, 2012.

[24] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8984–8994, June 2024.

[25] Hainan Ren, Li Lin, Chun-Hao Liu, Xin Wang, and Shu Hu. Improving generalization for ai-synthesized voice detection. In AAAI, 2025.

[26] Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In CVPR, 2025.

[27] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Li Hao, Jiaxin Ai, Qin Zou, Chen Li, and Zhongyuan Wang. Stacking brick by brick: Aligned feature isolation for incremental face forgery detection. In CVPR, 2025.

[28] Monk skin tone scale. In https://en.wikipedia.org/wiki/Monk_Skin_Tone_Scale. Wikipedia, The Free Encyclopedia.

[29] 1st workshop on new trends in ai-generated media and security. In https://xianhjiang.github.io/AIMS-24.

[30] Reddit. In https://www.reddit.com/.

[31] Stack exchange communities. In https://stackexchange.com/sites.

[32] Researchgate. In https://www.researchgate.net/topic/Artificial-Intelligence.

[33] Machine learning news google groups. In https://groups.google.com/g/ml-news.

[34] National society of black engineers. In https://www.nsbe.org/.

[35] Society of women engineers. In https://swe.org/.

[36] Association for women in science. In https://awis.org/.

## A Biography of all team members

**Prof. Shu Hu** is an assistant professor in the Department of Computer and Information Technology, Purdue University. He received the PhD degree in computer science and engineering from University at Buffalo, SUNY, in 2022. He was a postdoc at Carnegie Mellon University from 2022 to 2023. His research interests include machine learning, multimedia forensics, and computer vision. He is a member of IEEE. ***(Role: Coordinators; Data Providers; Baseline Method Providers)***

**Prof. Xin Wang** (*Senior Member, IEEE*) is an assistant professor with the Department of Epidemiology and Biostatistics, School of Public Health, University at Albany, SUNY. He received the PhD degree in computer science from the University at Albany, State University of New York (SUNY), in 2015. His research interests include artificial intelligence, reinforcement learning, deep learning, and their applications. ***(Role: Platform Administrators; Beta Testers)***

**Prof. Daniel S. Schiff** is an Assistant Professor of Technology Policy at Purdue University's Department of Political Science and the co-director of GRAIL, the Governance and Responsible AI Lab. Dr. Schiff studied Philosophy at Princeton University, focusing on robotics and intelligent systems, before completing a Master's in Social Policy at the University of Pennsylvania and PhD in Public Policy from the Georgia Institute of Technology. ***(Role: Data Providers)***

**Prof. Sachi Nandan Mohanty** (*Senior Member, IEEE*) is a professor in School of Computer Science and Engineering at VIT-AP University. He received the Ph.D. degree from Indian Institute of Technology Kharagpur, India, in 2015. He has authored/edited 42 books, published by IEEE-Wiley, Springer, Wiley, CRC Press, NOVA, and DeGruyter. He has published 241 international journals of international repute. His research interests include data mining, brain-computer interface, cognition, and computational intelligence. He was recognized as Top 2% World Scientists Ranking by Stanford University and Elsevier for years 2022, 2023, and 2024. ***(Role: Publication Chair)***

**Mr. Ryan Ofman** is the Head of AI Research at Deep Media AI. He received his degree from Yale University, where he focused on Machine Learning applications in Astrophysics. His research includes work on applying AI classification methods to galaxy morphology. His current research interests include deepfake detection, multimedia forensics, and promoting equitable use of AI tools. 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 ***(Role: Evaluators, Sponsor)***

**Mr. Narcis Bejtic** is a COO at Originality.AI. He graduated from University of Western Ontario with BA and MA degrees in political science. ***(Role: Sponsor)***

**Mr. Jon Gillham** is a Founder / CEO of Originality.AI. He has been involved in the SEO and Content Marketing world for over a decade. His career started with a portfolio of content sites, recently he sold 2 content marketing agencies, and he is the Co-Founder of MotionInvest.com, the leading place to buy and sell content websites. Through these experiences, he understands what web publishers need when it comes to verifying content is original. He is not For or Against AI content, he thinks it has a place in everyone's content strategy. However, he believes you as the publisher should be the one making the decision on when to use AI content. His Originality checking tool has been built with serious web publishers in mind! ***(Role: Sponsor)***

**Prof. Wenbin Zhang** Wenbin Zhang is an Assistant Professor in the Knight Foundation School of Computing & Information Sciences at Florida International University, and an Associate Member at the Te Ipu o te Mahara Artificial Intelligence Institute. His research investigates the theoretical foundations of machine learning with a focus on societal impact and welfare. In addition, he has worked in a number of application areas, highlighted by work on healthcare, digital forensics, geophysics, energy, transportation, forestry, and finance. He is a recipient of best paper awards/candidates at FAccT'23, ICDM'23, DAMI, and ICDM'21, as well as the NSF CRII Award and recognition in the AAAI'24 New Faculty Highlights. He also regularly serves in the organizing committees across computer science and interdisciplinary venues, most recently Travel Award Chair at AAAI'24, Volunteer Chair at WSDM'24 and Student Program Chair at AIES'23. ***(Role: Platform Administrators; Beta Testers)***

**Prof. Baoyuan Wu** (*Senior Member, IEEE*) is a tenured Associate Professor, and Assistant Dean (research) of School of Data Science, the Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen). He received the PhD degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2014. His research interests are 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. ***(Role: Beta Testers; Evaluators)***

**Dr. Cristian Canton** is the head of Responsible AI (RAI) at Meta, where he supports multiple organizations related to AI robustness, fairness, transparency, and legitimacy. Recently, have expanded his scope to cover all aspects of RAI for Generative AI. ***(Role: Evaluators)***

**Prof. Xiaoming Liu** (*Fellow of IEEE and IAPR*) is a Anil K. and Nandita Jain Endowed Professor and MSU Foundation Professor at the Department of Computer Science and Engineering of Michigan State University. He received the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University in 2004. Before joining MSU in Fall 2012, he was a research scientist at General Electric (GE) Global Research. His research interests include computer vision, machine learning, and biometrics. 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. ***(Role: Evaluators)***

**Prof. Luisa Verdoliva** (*Fellow of IEEE*) is a professor with the Department of Electrical Engineering and Information Technology, University Federico II, Naples, Italy. She is currently the **Editor-in-Chief** for IEEE Transactions on Information Forensics and Security (2025-). She is the recipient of a Google Faculty Research Award for Machine Perception (2018) and a TUM-IAS Hans Fischer Senior Fellowship (2020–2024). She was chair of the IFS TC (2021–2022). Her scientific interests are in the field of image and video processing, with main contributions in the area of multimedia forensics. ***(Role: Evaluators)***

**Prof. Siwei Lyu** (*Fellow of IEEE, IAPR, and AAIA, Distinguished Member of ACM*) is currently a SUNY Empire Innovation Professor with the Department of Computer Science and Engineering, the Director of the UB Media Forensic Lab (UB MDFL), and the founding Co-Director of the Center for Information Integrity (CII) with the University at Buffalo, State University of New York, Buffalo, NY, USA. Siwei's research interests include digital media forensics, computer vision, and machine learning. He is also a Senior Member of the Sigma Xi Society and a Member of the Omicron Delta Kappa Society. ***(Role: Evaluators)***