Iterative Binary Training

Vitor da Silva, Rosana Tomás, Fernando Cores and Francesc Giné
Department of Computer Engineering and Digital Design
University of Lleida
C/Jaume II, 69, 25001, Lleida, Spain
{vitor.dasilva, rosana.tomas, fernando.cores, francesc.gine}@udl.cat

Abstract

In this work, we present Iterative Binary Training, an effective training strategy designed to improve face anti-spoofing systems, especially when dealing with imbalanced datasets. Instead of treating all spoofing attacks at once, our method starts by training a binary classifier to distinguish bonafide faces from the most frequent spoofing type. Then, step by step, it adds new spoofing classes—moving from the most common to the rarest— and thus, the model gradually learns to handle a wider variety of attacks.

This method encourages the model to first focus on dominant spoofing patterns and later adapt to more challenging, less frequent attacks, reducing overfitting and improving generalization. We tested it using four deep learning models, including ViT-B/16, ViT-B/32, ResNeXt-101, and ResNet-50. All showed good performance with our method, with ResNeXt-101 standing out as the top performer.

Our approach does not rely on extra data, additional modalities, or ensembling techniques. Instead, it builds on standard tools like class-balanced loss functions and pre-trained backbones, making it easy to reproduce and deploy. The results suggest that Iterative Binary Training offers a promising direction for enhancing FAS systems in real-world scenarios.

1. Introduction

Face anti-spoofing (FAS) remains a dynamic and challenging field, especially as attackers continuously develop more sophisticated and varied attack strategies. The increasing diversity and realism of presentation attacks—ranging from high-resolution prints and video replays to 3D masks and adversarial digital forgeries—have pushed FAS systems to their limits.

In this context, the ICCV 2025 6th Face Anti-Spoofing Challenge [1] introduces a significant step forward: the release of the UniAttackData+ dataset [12], designed to reflect real-world complexities by combining multiple attack types and unseen scenarios within a unified benchmark. This dataset addresses critical limitations observed in earlier benchmarks, like CASIA-SURF [19] or CASIA-SURF CeFA [9], which often relied on narrow acquisition settings, a limited number of subjects, or incomplete coverage of spoofing modalities.

Moreover, recent works [5, 7, 11] have emphasized the need to unify physical and digital attack detection. Fang et al. [5] introduced a vision-language model capable of identifying diverse attack types within a single framework, whereas He et al. [7] proposed simulated data augmentation via SPSC and SDSC to expose models to unseen spoofing clues. Liu et al. [11] proposed Class-Free Prompt Learning (CFPL), a CLIP-based method that improves generalization by dynamically adapting classifier weights using style- and content-aware prompts, without relying on fixed spoof classes or domain labels. These directions point toward training paradigms that prioritize generalization over fitting known spoof taxonomies.

Unlike previous editions, which focused predominantly on performance metrics, the ICCV 2025 6th Face Anti-Spoofing Challenge encourages the exploration of innovative and practical solutions that go beyond accuracy scores. According to the organizers, the goal of this year is to foster the development of robust and generalizable approaches, with an emphasis on the design of methodologies and the understanding of underlying limitations. This shift creates a unique opportunity to rethink how the training and evaluation of FAS models can be faced.

Motivated by this change, we propose to investigate not only the architecture or performance of models, but the way in which data is presented to the classifier during training. Our approach is grounded on two core assumptions: (1) the nature of the FAS problem can be reformulated as a binary classification task, where the model learns to distinguish between bonafide and spoofed samples, regardless of spoofing modality [7], and (2) FAS datasets are inherently imbalanced, with bonafide samples often underrepresented or overly homogeneous compared to the diversity of

attacks [9].

To address these challenges, we introduce Iterative Binary Training (ITB), a data exposure strategy that structures learning as a progressive sequence of binary tasks. Starting with the most frequent spoofing class and incrementally adding rarer ones, this method encourages early generalization while systematically reducing class bias and improving resilience against unseen attacks.

The remainder of this paper is structured as follows. Section 2 reviews the state-of-the-art in face anti-spoofing (FAS) datasets and training strategies. Section 3 introduces the proposed Iterative Binary Training (IBT) method, a progressive strategy tailored to imbalanced classification tasks in the FAS context. Section 4 details the experimental setup and evaluates IBT using the UniAttackData+ benchmark, comparing its performance with standard training and across four different model architectures: ViT-B/16, ViT-B/32, ResNeXt-101, and ResNet-50. Section 5 discusses the results, highlighting the strengths and limitations of IBT. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

Towards Unified Detection of Physical and Digital Attacks. While most prior work treated Physical Attack Detection (PAD) and Digital Attack Detection (DAD) as separate tasks, recent efforts advocate for a unified treatment. Fang *et al*. [5] proposed *UniAttackDetection*, a vision-language model trained on the *UniAttackData* dataset, which includes ID-consistent samples across 2 physical and 12 digital attack types for over 1,800 subjects. Their method introduces prompt-based modules — Teacher-Student Prompts, Unified Knowledge Mining, and Sample-Level Prompt Interaction — that enable simultaneous learning of unified and specific knowledge representations.

He *et al*. [7] advanced this by introducing simulated spoofing clues to bonafide data, bridging the physical-digital domain gap during training. Their approach uses synthetic transformations (SPSC and SDSC) to enhance the model's robustness to unseen attack types, achieving State-of-the-Art (SoA) results on cross-modal and cross-dataset settings. These two works redefine how FAS systems can generalize across modalities.

Benchmark Evolution and Dataset Scope. Dataset diversity is central to robust generalization in FAS. Zhang et al. introduced the CASIA-SURF dataset [19], a large-scale multi-modal benchmark that addresses limitations of scale and modality in earlier datasets. It includes RGB, depth, and IR modalities across more than 20,000 video clips. Later, CASIA-SURF CeFA [9] further expanded coverage to multi-ethnic populations and provided a strong foundation for cross-ethnicity evaluation. Liu *et al*. [10] extended this with contrastive learning strategies to improve mask PAD under 3D conditions.

Building on these efforts, UniAttackData [5] introduced a unified benchmark combining identity-consistent samples across diverse spoof types. It enabled the creation of protocols for unseen attack evaluation, which are essential for assessing generalization. More recently, UniAttackData+ [12] extended this benchmark by incorporating additional spoof categories and updated evaluation protocols, and currently represents the most comprehensive physical-digital dataset in the field. In contrast to earlier benchmarks, which often merged separate datasets without consistent identity alignment, both UniAttackData and its extended version ensure that each subject includes all spoof variants, minimizing label leakage and enhancing model robustness.

Learning Strategies in Face Anti-Spoofing. Conventional supervised learning has laid the foundation for many FAS systems. Typically, deep neural networks are trained to classify bonafide versus spoof samples using standard losses like cross-entropy. While this setup yields competitive results in closed-set settings, models often fail to generalize to unseen attack types due to overfitting and dataset bias.

To overcome these limitations, researchers have explored alternative training paradigms such as meta-learning and multi-task learning. These strategies aim to improve adaptability and robustness by exposing models to auxiliary tasks and simulated domain shifts. For instance, Chuang *et al*. [3] proposed a unified model that integrates face parsing, depth estimation, and spoof classification using a one-side triplet loss. This approach helps the model focus on live features and improves generalization across spoof types and domains.

Another direction is one-class learning and anomaly detection [8], where models are trained only on bonafide samples and are expected to reject anything deviating from the learned distribution. These models tend to generalize better to unknown attacks, although at the cost of reduced sensitivity.

Building on these foundations, curriculum learning has emerged as a promising approach. Initially introduced by Bengio *et al*. [2], this paradigm organizes the training process by presenting easier examples first and gradually increasing difficulty. While curriculum learning has shown success in general computer vision tasks [17], its application to FAS remains relatively limited.

Some efforts have adopted curriculum-like approaches. Quan *et al*. [15] proposed a progressive transfer learning framework that starts from a small set of labeled data and in-

crementally adds high-confidence pseudo-labeled samples. Their method leverages temporal consistency and adaptive selection to expose the model to gradually more complex spoofing scenarios, yielding strong performance under domain shifts.

In this context, our work introduces two innovations. First, we explicitly frame FAS as a binary classification problem, treating bonafide as a fixed anchor class while spoofing types vary across training stages. Second, we propose a reverse curriculum learning strategy, where the model is first exposed to the most frequent (and typically easiest) spoof class and progressively learns to distinguish rarer and more challenging attacks. This learning schedule enables the model to form strong early representations and improves its robustness to less common spoof types.

To our knowledge, this combination—binary task formulation with an inverse curriculum schedule—has not been explicitly explored in the FAS literature. It offers a promising new direction for improving generalization in highly imbalanced and heterogeneous anti-spoofing scenarios.

3. Iterative Binary Training Method

We propose an Iterative Binary Training (IBT) strategy for classification tasks in imbalanced datasets, specifically designed for face anti-spoofing. The method assumes a fixed bonafide class and introduces spoofing classes progressively across multiple training phases.

The motivation for using binary classification is to encourage the model to generalize to *unseen and unknown spoofing attacks*, learning to recognize bonafide samples regardless of the spoofing type presented.

While traditional curriculum learning [17] typically progresses from easy to hard samples, our method follows a *reverse class frequency schedule*: training begins with the most frequent spoofing class and progressively incorporates less frequent classes in each new phase.

At each iteration:

- The model is trained as a binary classifier: bonafide vs. a single spoofing class.
- After convergence or a fixed number of epochs, a new spoofing class is introduced, and training continues.

Algorithm 1 summarizes this procedure in pseudo-code:

Algorithm 1 Iterative Binary Training (IBT)

Require: Live samples $\mathcal{D}_{\text{live}}$; spoofing classe $\mathcal{D}_{\text{spoof}}$; spoofing subsets $\mathcal{D}_{c_i}$ (rarest $\rightarrow$ frequent); backbone Φ_0; epochs per loop E
Ensure: Trained classifier $\Phi^\star$

$\Phi \leftarrow \Phi_0$ ▷ binary head (live / spoof)
for all spoofing class $\mathcal{D}_{c_i}$ in $\mathcal{D}_{\text{spoof}}$ **do**
  $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{live}} \cup \mathcal{D}_{c_i}$
  for $e = 1$ **to** E **do**
    Train Φ on $\mathcal{D}_{\text{train}}$
  end for
end for
return $\Phi^\star \leftarrow \Phi$

- Φ is the convolutional backbone being fine-tuned; it is initialized from a pretrained model Φ_0 with a two-class (live vs. spoof) faces.
- $\mathcal{D}_{\text{live}}$ contains all bona fide images, and $\mathcal{D}_{\text{spoof}}$ contains all spoofing classes ($\mathcal{D}_{\text{spoof}} = \cup_{i=1}^{N} \mathcal{D}_{c_i}$), where $\mathcal{D}_{c_i}$ denotes the set of samples belonging to the i-th spoofing class c_i, and N is the total number of spoofing classes.
- Spoofing classes are introduced *iteratively*, from the rarest to the most frequent. After adding a class, Φ is retrained for E epochs on the binary set $(\mathcal{D}_{\text{live}} \cup \mathcal{D}_{c_i})$.

This method facilitates early learning from dominant patterns, incrementally increases robustness by exposing the model to more diverse and rare attacks, and reduces both class bias and catastrophic forgetting. Furthermore, it integrates well with transfer learning and data augmentation, without requiring full dataset rebalancing at the start.

4. Experiments

4.1. Experimental Setup

AttackData+ Dataset. We conduct all our experiments using the UniAttackData+ dataset, officially released as part of the ICCV 2025 6th Face Anti-Spoofing Challenge [11, 12]. This benchmark extends the original UniAttackData [5] to include a broader set of spoofing types, covering both physical and digital modalities. The dataset contains diverse attack samples and subjects of multiple ethnic backgrounds. Spoofing classes are organized hierarchically by modality and subcategory. Specifically, physical attacks (`1_`) include 2D types such as *Print*, *Replay*, and *Cutouts*, and 3D types such as *Transparent*, *Plaster*, and *Resin*. Digital attacks (`2_`) include three major categories: *Digital Manipulation* (e.g., Face-Swap, Attribute-Edit), *Digital Adversarial* (e.g., Pixel-Level, Semantic-Level), and *Digital Generation* (e.g., ID-Consistent, Style, Prompt-based).

Following the official protocol of the challenge, the training and validation sets are disjoint and contain mutually exclusive attack samples. Table 1 summarizes the number of instances per class in both sets.

Protocol	Type	Class	Label ID	#Samples	#Total Samples
Train	Live	Live Face	0_0_0	839	22367
	Physical	Print	1_0_0	43	
	Physical	Replay	1_0_1	109	
	Physical	Cutouts	1_0_2	79	
	Digital	Face-Swap	2_0_1	6160	
	Digital	Attribute-Edit	2_0_0	1476	
	Digital	Video-Driven	2_0_2	1540	
	Digital	Pixel-Level	2_1_0	8364	
	Digital	Semantic-Level	2_1_1	3757	
Eval	Live	Live Face	0_0_0	13	5396
	Physical	Print	1_0_0	12	
	Physical	Replay	1_0_1	26	
	Physical	Cutouts	1_0_2	20	
	Digital	Face-Swap	2_0_1	1555	
	Digital	Attribute-Edit	2_0_0	385	
	Digital	Video-Driven	2_0_2	354	
	Digital	Pixel-Level	2_1_0	2053	
	Digital	Semantic-Level	2_1_1	978	
Test	*The test protocol remains unpublished and is solely used by the challenge organizers for official scoring.*				9083

Table 1. Distribution of samples in the UniAttackData+ dataset across training, evaluation, and test protocols.

Models. Four backbone models were tested: ViT-B/16 and ViT-B/32 [4], ResNeXt-101 [18], and ResNet-50 [6]. ViT-B/16 and ViT-B/32 are Vision Transformers that model image patches as tokens and rely entirely on self-attention, differing mainly in patch size and resolution. ResNet-50 is a standard convolutional architecture with residual connections, while ResNeXt-101 is a more powerful CNN that extends ResNet by introducing group convolutions to increase representational capacity with controlled complexity. All models were evaluated under two initialization regimes: standard pretrained weights (ImageNet) and from scratch.

Training Configuration. All models were trained using the AdamW optimizer [13] with a learning rate of 1×10^{-4} and weight decay of 0.01. We resized input images to 224×224 and applied no augmentations. The training batch size was set to 32, and each binary phase was trained for 30 epochs. Cross-entropy loss [16] was used, with class-balanced weights computed using the Scikit-learn [14] `compute_class_weight` function.

Evaluation Metrics. The evaluation protocol used to assess model performance follows standard practices in the field of Face Anti-Spoofing (FAS). In particular, we adopt three widely accepted metrics:

- **Attack Presentation Classification Error Rate (APCER)**: It measures the proportion of attack samples incorrectly classified as bonafide.
- **Bona Fide Presentation Classification Error Rate (BPCER)**: It measures the proportion of bonafide samples incorrectly classified as attacks.
- **Average Classification Error Rate (ACER)**: This is the average of APCER and BPCER.

These metrics are formally defined as:

$$\begin{aligned} \text{APCER} &= \frac{FP}{FP+TN} \\ \text{BPCER} &= \frac{FN}{FN+TP} \\ \text{ACER} &= \frac{\text{APCER}+\text{BPCER}}{2}, \end{aligned} \tag{1}$$

where FP, FN, TP, and TN denote the number of false positives, false negatives, true positives, and true negatives, respectively. The ACER score is used to determine final rankings in the ICCV 2025 6th Face Anti-Spoofing Challenge [1] and is the metric used in our experimentation.

4.2. Comparison with Baseline

For comparison, we define a baseline *standard training* (STD) strategy in which the model is trained in a single stage using all spoofing classes simultaneously, without any curriculum or iterative exposure. Figure 1 presents the ACER evolution across epochs (5 to 30) for four different models, comparing the baseline STD with our proposed

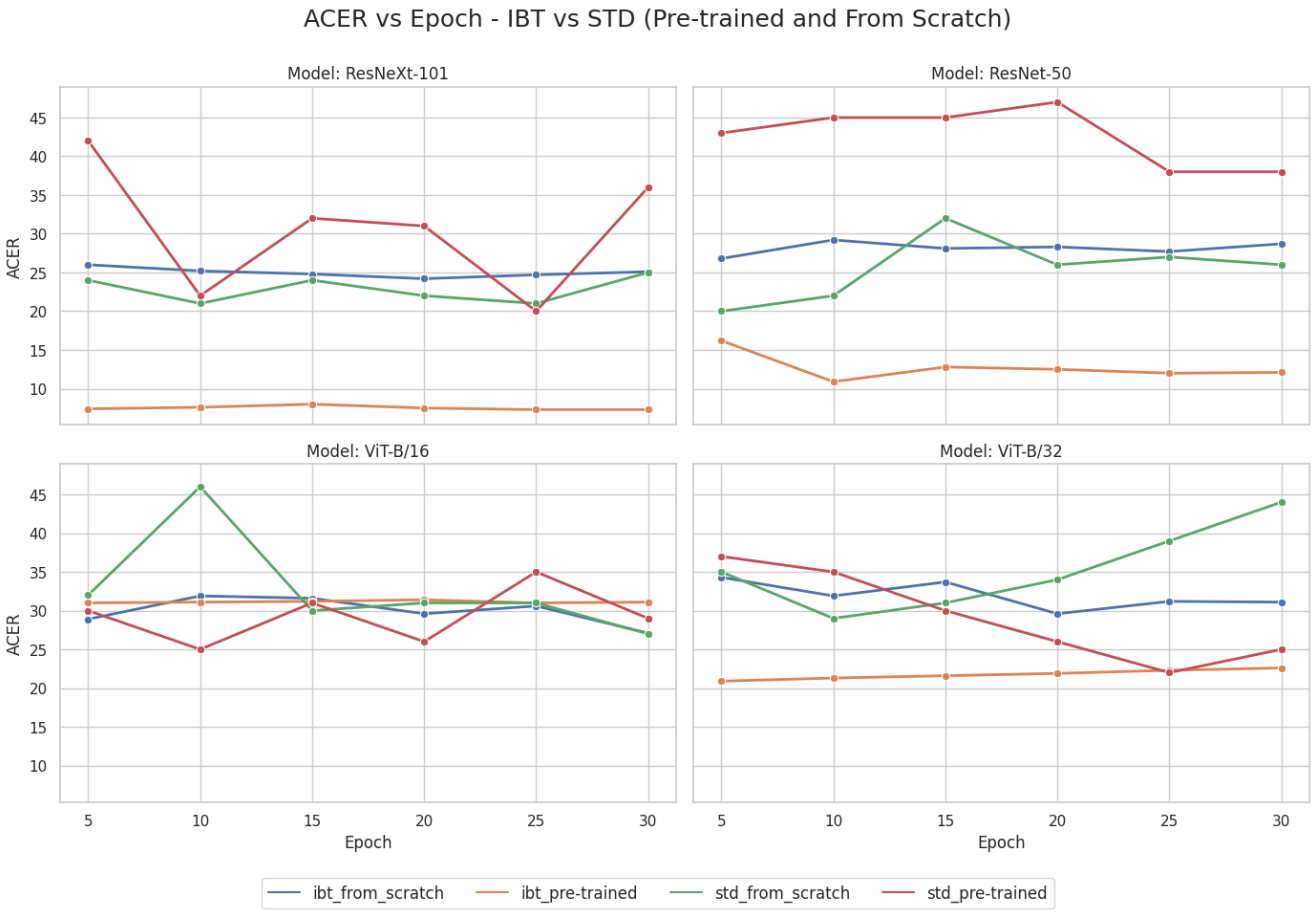

Figure 1. IBT vs STD per epoch

Model	BEST ACER (%) - From Scratch		BEST ACER (%) - Pre-Trained	
	Standard Training	Iterative Binary Training	Standard Training	Iterative Binary Training
ResNeXt-101	21.0	24.2	20.0	**7.3**
ResNet-50	20.0	26.8	38.0	**10.9**
ViT-B/16	27.0	27.1	**25.0**	31.0
ViT-B/32	29.0	29.6	22.0	**20.9**

Table 2. Best ACER comparison across models trained from scratch and with pre-trained weights, using standard and iterative binary training.

IBT, both with and without pre-trained initialization.

The two convolutional-based models, ResNeXt-101 and ResNet-50, show substantial improvements when using IBT in conjunction with pretrained weights. In particular, ResNeXt-101 achieves the best overall performance, reducing its ACER from 20.0% (standard training) to 7.3% with IBT. Similarly, ResNet-50 improves from 38.0% to 10.9%, highlighting the effectiveness of combining iterative learning with prior knowledge from large-scale pretraining.

In contrast, the transformer-based models demonstrate differing patterns of behavior. While ViT-B/16 shows minimal variation across configurations, with similar ACER values under all training strategies, ViT-B/32 consistently benefits from IBT — especially when using pretrained initialization — confirming its sensitivity to data exposure strategies.

When comparing models trained from scratch, the performance gap between STD and IBT is narrower. In this setting, convolutional models slightly favor standard training, while transformer models tend to benefit more from the

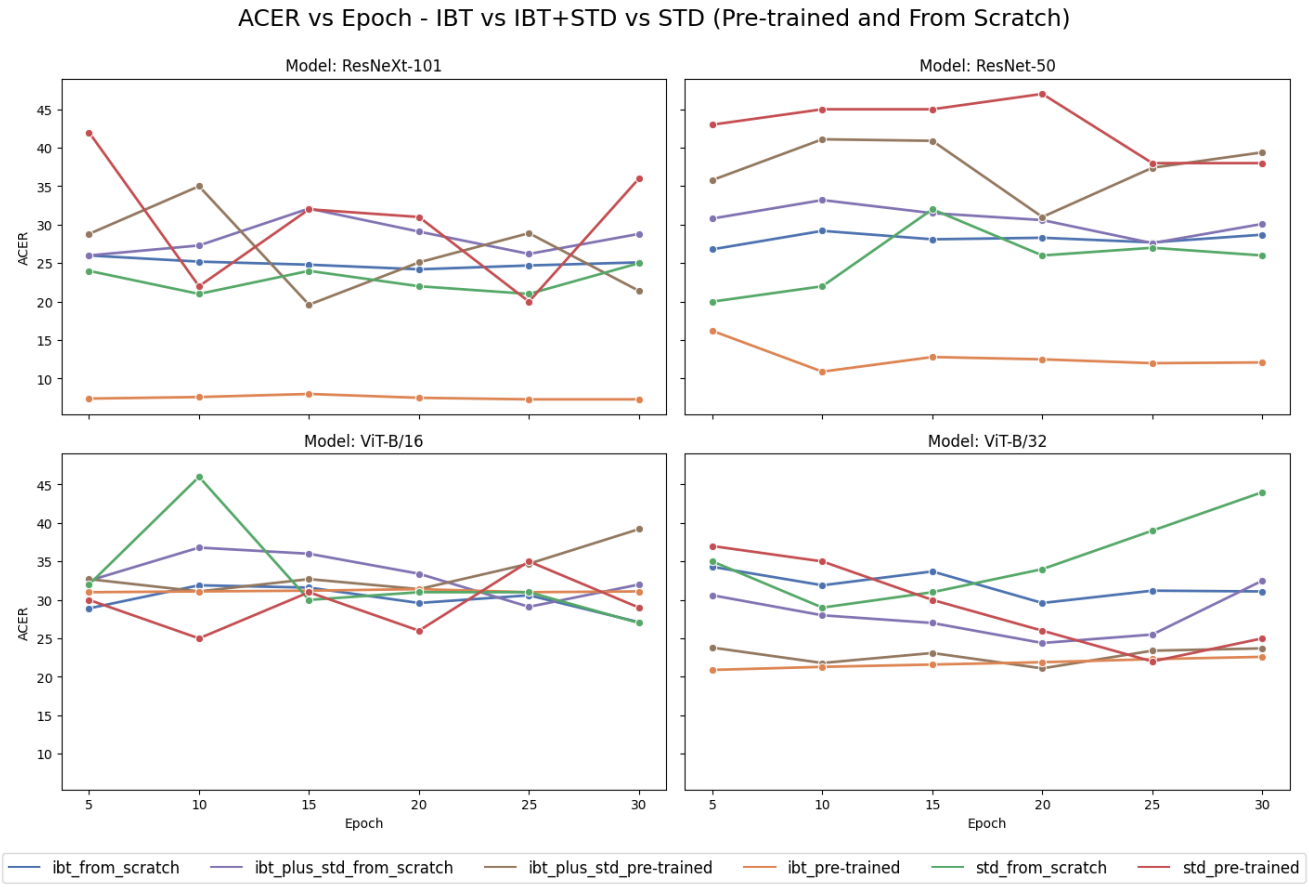

Figure 2. IBT vs IBT+STD vs STD per epoch (Pre-trained and From Scratch)

Model	BEST ACER (%) - From Scratch			BEST ACER (%) - Pre-Trained		
	STD	IBT	IBT+STD	STD	IBT	IBT+STD
ResNeXt-101	21.0	24.2	26.0	20.0	**7.3**	19.6
ResNet-50	20.0	26.8	27.6	38.0	**10.9**	31.0
ViT-B/16	27.0	27.1	29.1	**25.0**	31.0	31.1
ViT-B/32	29.0	29.6	24.4	22.0	**20.9**	21.1

Table 3. Best ACER comparison - Ablation study IBT+STD

IBT strategy.

Overall, IBT method demonstrates notable improvements in ACER for most architectures, particularly when pretrained weights are available. These results validate our hypothesis that structured, progressive exposure to spoof types enhances generalization. Table 2 summarizes the best ACER scores obtained for each configuration.

4.3. Ablation Study

IBT followed by Standard Training. This experiment was designed to evaluate whether the Iterative Binary Training (IBT) strategy can serve as an effective pre-training phase. To test this, we appended a final standard training step (STD) after the last binary iteration, where the model is retrained using all classes together. The goal was to assess whether combining the binary-focused learning with standard multi-class training would improve generalization.

The results shown in Figure 2 demonstrate that IBT followed by Standard Training (IBT+STD) consistently outperforms standard training (STD) across most epochs, while still performing worse than the pure IBT strategy.

Table 3 reports the best ACER scores obtained from models trained from scratch and using pre-trained weights across STD, IBT, and the combined IBT+STD setup.

Occlusion Type	Iter 1	Iter 2	Iter 3	Iter 4	Iter 5	All (STD)
Physical Occlusion (PO)	Pixel-Level	Face-Swap	Semantic-Level	Video-Driven	Attribute-Edit	+ All Physical
Digital Occlusion (DO)	Replay	Cutouts	Print	–	–	+ All Digital

Table 4. Occlusion-based ablation protocol per iteration, with PO using only digital attacks and DO using only physical attacks.

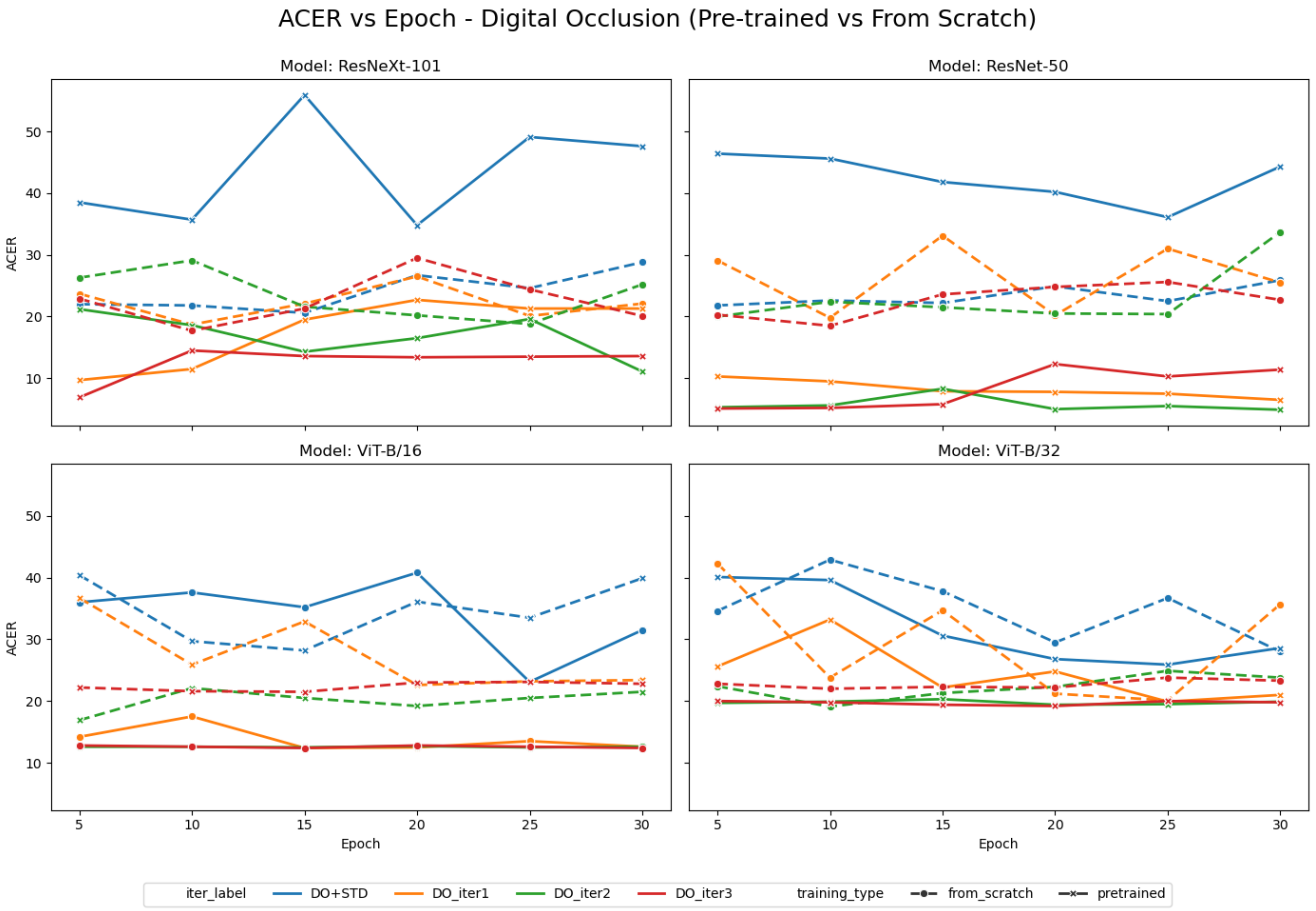

Figure 3. Digital Occlusion per Epoch

Model	BEST ACER (%) - Pre-Trained						
	STD	IBT	IBT+STD	PO	PO+STD	DO	DO+STD
ResNeXt-101	20.0	7.3	19.6	25.8	21.6	**6.9**	34.8
ResNet-50	38.0	10.9	31.0	32.4	33.4	**5.1**	36.1
ViT-B/16	25.0	31.0	31.1	26.9	27.9	**12.4**	23.1
ViT-B/32	22.0	20.9	21.1	24.1	21.4	**19.2**	25.9

Table 5. Best ACER (%) for pre-trained models across standard training, Iterative Binary Training (IBT), and occlusion-based configurations. PO refers to Physical Occlusion and DO to Digital Occlusion, both part of the ablation study.

Occlusion-based IBT: Physical vs. Digital Attacks. To further investigate the impact of different types of spoofing attacks on training performance, we conducted a targeted ablation study using occlusion. We trained models using IBT method while systematically excluding all physical attack classes (PO – Physical Occlusion) and excluding all

digital attack classes (DO – Digital Occlusion) from the binary training steps. Table 4 outlines the occlusion protocol, detailing which spoofing classes were included in the binary training at each iteration for both PO and DO settings. Classes were introduced in descending order of frequency, and all configurations included live samples. A final standard training step (denoted as "STD") reintroduced all classes to assess recovery in performance. These experiments aim to evaluate the relative importance of each attack group in the learning process, and their contribution to generalization when reintroduced via standard training (PO+STD and DO+STD).

Table 5 shows the best ACER scores for each occlusion configuration using only pre-trained models, since from-scratch models consistently performed worse with ACER above 20%.

The results show that Digital Occlusion (DO) not only preserves competitive performance but in fact outperforms standard training (STD) and Iterative Binary Training (IBT) without occlusion across all models. As shown in Table 5, DO achieves the lowest ACER for every architecture, highlighting its effectiveness as a training strategy. The full evolution of ACER scores across training epochs and occlusion iterations — comparing from scratch and pretrained models — is illustrated in Figure 3. This is especially remarkable in the case of ResNet-50 and ResNeXt-101, where ACER drops from 38.0% to 5.1% and from 20.0% to 6.9%, respectively. These improvements suggest that training with only physical attacks forces the model to generalize better to digital attacks, which appear more diverse and harder to classify. Surprisingly, adding a final standard training step (DO+STD) does not improve performance and in some cases even degrades it in comparison with STD, possibly due to loss of the specialized representation learned during the binary-focused training. In contrast, the Physical Occlusion (PO) strategy leads to significantly worse results very close to the standard training. These findings reinforce the idea that physical attacks play a key role in effective model generalization, and that carefully selected occlusion strategies can outperform traditional training methods.

5. Discussion

While Iterative Binary Training has proven to be a stable and effective strategy for handling class imbalance in face anti-spoofing, there are several directions worth exploring to further improve its performance and generalization capabilities.

First, although our current scheduling strategy is based on spoof class frequency (i.e., starting from the most frequent attack), it may be beneficial to experiment with alternative curricula. For example, ordering classes based on difficulty, visual similarity, or domain shift could allow for more targeted knowledge transfer across spoof types. Similarly, grouping spoof classes by physical modality (e.g., print, replay, 3D) might better align the curriculum with the underlying structure of the data.

Second, the binary formulation, while simplifying the training dynamics, might limit the model's ability to simultaneously learn shared spoof features. Incorporating multi-label soft supervision or intermediate knowledge distillation could allow the model to retain useful spoof-specific representations while preserving the benefits of the binary focus.

Third, although our method does not require additional data or architectural changes, its iterative nature increases training time linearly with the number of spoofing classes. This motivates the investigation of adaptive iteration schedules, where early stopping or dynamic class inclusion could reduce computational cost while maintaining robustness.

Finally, we note that our current experiments are constrained to a single dataset (UniAttackData+). To truly evaluate the generalization ability of this method, future work should explore cross-dataset experiments and domain adaptation setups, where class imbalance and unseen attack types are even more pronounced.

In summary, Iterative Binary Training opens up a opportunity for future research on data exposure strategies in FAS, with significant potential to be extended through more adaptive, structured, and multi-task training components.

6. Conclusions

In this work, we introduced Iterative Binary Training, a simple yet effective strategy designed to address data imbalance and generalization challenges in face anti-spoofing. Despite not achieving top-3 performance in the ICCV 2025 competition, our approach delivered satisfactory results and ranked among the top 8 teams, demonstrating the practical potential of restructuring the training process through iterative, frequency-based data exposure.

Given that the full dataset from the competition (UniAttackData+) has not yet been publicly released, our evaluation was constrained to the test subset accessed exclusively through the competition's official scoring system. Still, our method consistently led to stable training and correct convergence across several architectures, particularly ResNeXt-101.

These initial findings suggest that Iterative Binary Training can offer meaningful improvements in learning dynamics without modifying model architectures or requiring additional data. Nevertheless, further studies are needed to validate this approach across diverse datasets and under alternative exposure strategies—including different curriculum orders, grouping schemes, or domain-specific data augmentation.

Acknowledgements

This work has been granted by the Ministerio de Ciencia e Innovación MCIN AEI/10.13039/501100011033 under contract PID2023-146193OB-I00.

We would also like to thank the organizers of the 6th Face Anti-Spoofing Challenge Workshop for their support during the evaluation process. In particular, we are grateful to Ajian Liu for his assistance in executing the official scoring program, which contributed to improving the experimental results presented in this paper.

References

[1] 6th face anti-spoofing challenge (iccv 2025 workshop). https://sites.google.com/view/face-anti-spoofing-challenge/welcome/challengeiccv2025, 2025. Access: 10/07/2025. 1, 4

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2

[3] Chu-Chun Chuang, Chien-Yi Wang, and Shang-Hong Lai. Generalized face anti-spoofing via multi-task learning and one-side meta triplet loss. In *2023 IEEE 17th international conference on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2023. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 4

[5] Hao Fang, Ajian Liu, Haocheng Yuan, Junze Zheng, Dingheng Zeng, Yanhong Liu, Jiankang Deng, Sergio Escalera, Xiaoming Liu, Jun Wan, et al. Unified physical-digital face attack detection. *arXiv preprint arXiv:2401.17699*, 2024. 1, 2, 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[7] Xianhua He, Dashuang Liang, Song Yang, Zhanlong Hao, Hui Ma, Binjie Mao, Xi Li, Yao Wang, Pengfei Yan, and Ajian Liu. Joint physical-digital facial attack detection via simulating spoofing clues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 995–1004, 2024. 1, 2

[8] Pei-Kai Huang, Cheng-Hsuan Chiang, Tzu-Hsien Chen, Jun-Xiong Chong, Tyng-Luh Liu, and Chiou-Ting Hsu. One-class face anti-spoofing via spoof cue map-guided feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 277–286, 2024. 2

[9] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1179–1187, 2021. 1, 2

[10] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE transactions on information forensics and security*, 17: 2497–2507, 2022. 2

[11] Ajian Liu, Shuai Xue, Jianwen Gan, Jun Wan, Yanyan Liang, Jiankang Deng, Sergio Escalera, and Zhen Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 3

[12] Ajian Liu, Haocheng Yuan, Xiao Guo, Hui Ma, Wanyi Zhuang, Changtao Miao, Yan Hong, Chuanbiao Song, Jun Lan, Qi Chu, et al. Benchmarking unified face attack detection via hierarchical prompt tuning. *arXiv preprint arXiv:2505.13327*, 2025. 1, 2, 3

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 4

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 4

[15] Ruijie Quan, Yu Wu, Xin Yu, and Yi Yang. Progressive transfer learning for face anti-spoofing. *IEEE Transactions on Image Processing*, 30:3946–3955, 2021. 2

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 4

[17] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2022. 2, 3

[18] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017. 4

[19] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. 1, 2