



This is an electronic reprint of the original article. This reprint may differ from the original in pagination and typographic detail.

# Muhammad, Usman; Laaksonen, Jorma; Romaissa Beddiar, Djamila; Oussalah, Mourad **Domain Generalization via Ensemble Stacking for Face Presentation Attack Detection**

Published in: International Journal of Computer Vision

*DOI:* 10.1007/s11263-024-02152-1

Published: 01/12/2024

Document Version Publisher's PDF, also known as Version of record

Published under the following license: CC BY

Please cite the original version:

Muhammad, U., Laaksonen, J., Romaissa Beddiar, D., & Oussalah, M. (2024). Domain Generalization via Ensemble Stacking for Face Presentation Attack Detection. *International Journal of Computer Vision*, *132*(12), 5759-5782. https://doi.org/10.1007/s11263-024-02152-1

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.



## Domain Generalization via Ensemble Stacking for Face Presentation Attack Detection

Usman Muhammad<sup>1,2</sup> · Jorma Laaksonen<sup>1</sup> · Djamila Romaissa Beddiar<sup>2</sup> · Mourad Oussalah<sup>2</sup>

Received: 15 September 2023 / Accepted: 13 June 2024 / Published online: 25 June 2024 © The Author(s) 2024

#### Abstract

Face presentation attack detection (PAD) plays a pivotal role in securing face recognition systems against spoofing attacks. Although great progress has been made in designing face PAD methods, developing a model that can generalize well to unseen test domains remains a significant challenge. Moreover, due to the different types of spoofing attacks, creating a dataset with a sufficient number of samples for training deep neural networks is a laborious task. This work proposes a comprehensive solution that combines synthetic data generation and deep ensemble learning to enhance the generalization capabilities of face PAD. Specifically, synthetic data is generated by blending a static image with spatiotemporal-encoded images using alpha composition and video distillation. In this way, we simulate motion blur with varying alpha values, thereby generating diverse subsets of synthetic data that contribute to a more enriched training set. Furthermore, multiple base models are trained on each subset of synthetic data using stacked ensemble learning. This allows the models to learn complementary features and representations from different synthetic subsets. The meta-features generated by the base models, leveraging their complementary information to better handle unseen target domains and enhance overall performance. Experimental results from seven datasets—WMCA, CASIA-SURF, OULU-NPU, CASIA-MFSD, Replay-Attack, MSU-MFSD, and SiW-Mv2—highlight the potential to enhance presentation attack detection by using large-scale synthetic data and a stacking-based ensemble approach.

Keywords Face anti-spoofing · Ensemble learning · Deep learning · Synthetic data · Recurrent neural network

### **1** Introduction

Over the past few decades, facial recognition (FR) technology has been frequently used in numerous real-world applications, such as mobile payments, access control, immigration, education, surveillance, and healthcare (Kim et al., 2022). The accuracy of FR is no longer a major concern, with the error rate dropping to 0.08%, according to the National Institute of Standards and Technology (NIST) (Grother et al., 2019). Despite its great success, a simple FR system might be vulnerable to spoofing, known as a presentation attack.

Communicated by Segio Escalera.

Usman Muhammad usman.muhammad@aalto.fi For instance, print attacks, video replays, and 3D masks are the most common attacks reported recently in the face antispoofing domain (Muhammad & Oussalah, 2022, 2023; Wu et al., 2021; Jia et al., 2020; Arashloo, 2020). To counter these attacks, several hand-crafted and deep representation methods have been proposed to protect FR systems (Boulkenafet et al., 2016; Liu et al., 2021; Shao et al., 2020; Muhammad et al., 2022a; Wang et al., 2020; Saha et al., 2020; Muhammad & Hadid, 2019; Shao et al., 2019). Many of these models have reported promising performance in the intra-domain testing scenario. However, their performances remain limited in the cross-dataset testing scenario due to the distributional discrepancy between the source domain and the target domain.

#### **1.1 Domain Adaptation and Generalization**

In the context of cross-dataset testing scenarios, a key contributing factor to the performance limitations of deep learning models can often be attributed to the restricted

<sup>&</sup>lt;sup>1</sup> Department of Computer Science, Aalto University, Espoo, Finland

<sup>&</sup>lt;sup>2</sup> Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland

size or inadequacy of the training dataset. Another reason is the inherent assumption in many face presentation attack detection methods that the training and testing data come from the same target distribution. This raises several questions. For instance, if a model is trained on cut photo attack images, would it work on mask attack images? What if a model is trained only on replay attack images but tested on warped photo attacks? Is it possible to deploy a model that is trained using different illumination conditions and background scenes under controlled lighting scenarios? Answers to all these questions depend on how a machine learning model can deal with such domain shift problems. Therefore, domain generalization (DG) is essential when the objective is to build a model that can perform well in entirely new and diverse domains that were not seen during its training. As shown in Fig. 1, domain generalization refers to the task of training a model on data from multiple source domains and then deploying it on a new, unseen target domain. To address the domain generalization issue, several face anti-spoofing methods, such as adversarial learning (Liu et al., 2022), metapattern learning (Cai et al., 2022), generative domain adaptation (Zhou et al., 2022), hypothesis verification (Liu et al., 2022), and cross-adversarial learning (Huang et al., 2022), have been introduced to improve the generalization ability of the model.

Since generalization is a fundamental challenge in machine learning, researchers have explored various generalizationrelated topics such as meta-learning (learning to learn), regularization techniques, ensemble learning, and data augmentation (Wang et al., 2020; Saha et al., 2020; Shao et al., 2019; Liu & Liu, 2022). In particular, domain generalization is important in face anti-spoofing because collecting and annotating new datasets in real-world scenarios can be expensive and time-consuming. On the other hand, domain adaptation (DA) approaches focus on adapting a model trained on one or multiple source domains to perform well on a specific target domain that is different from the source domains. The main difference between these two approaches is that domain adaptation assumes that the target domain is known during model training, while domain generalization does not. For instance, the research in Jia et al. (2020) relies on a shared feature space and assumes that it would also be invariant to domain shift. However, this assumption has limitations in face PAD, because as the source domains become more diverse, learning a domain-invariant model becomes more difficult (Zhou et al., 2021). Instead of concentrating on some domain-specific cues, such as paper texture, diverse data can help the model generalize better if more generalized cues are shared by all source domains (Shao et al., 2019). As spoofing attacks are launched physically by malicious hackers outside the control of the biometric system, domain generalization is generally more important than

domain adaptation because DG does not require the target domain to be known during training.

In most face PAD databases, both real and spoofed classes exhibit spatiotemporal attributes, offering valuable insights into facial movements, texture changes, and dynamic characteristics, which are essential for effectively distinguishing genuine faces from spoofing attacks. For instance, irregular patterns are often observed in spoofing attacks, particularly with textured materials like paper or fabric, and these patterns become more evident in video sequences. Similarly, print attacks can be detected by hand movements or material reflections, while replay attacks reveal artifacts caused by screen sloping. These spatiotemporal effects can help improve the generalization ability of the model. However, the presence of noisy camera movements poses a challenge that can make the detection of spoofing attacks more difficult. This is especially relevant in the context of face PAD, where the deep learning models need to effectively analyze various attributes, patterns, and cues to distinguish between genuine actions and spoofing attacks. Thus, effectively analyzing both spatial and temporal information is essential for improving the robustness of deep learning-based PAD models. This enables more accurate detection and prevention of spoofing attacks, ensuring higher security and reliability in face biometric authentication processes.

#### **1.2 Ensemble Methods and Their Limitations**

Recently, ensemble methods have demonstrated increased generalization capacity for unseen attacks. The concept behind ensemble learning involves aggregating the predictions of multiple models and combining them to produce one optimal predictive model. In particular, this combination of models expects to produce significantly improved results. For instance, by ensembling diverse face-spoofing attacks as source domains, researchers achieved the top-ranking position in the ChaLearn LAP face anti-spoofing challenge, showcasing the effectiveness of employing an ensemble technique (Parkin & Grinchuk, 2019). This achievement demonstrates the capacity of a diverse ensemble of models to enhance the stability and efficacy of a face anti-spoofing system. Additionally, an ensemble learning approach was used in Vareto and Schwartz (2021) with three different models, while a score fusion was applied to enhance the generalization of face PAD. Similarly, a depth-based ensemble learning approach was introduced in Jiang and Sun (2022), incorporating multiple domain-specific modules, to optimize spoof detection. In Fatemifar et al. (2019), the authors combined predictions from multiple models using a simple weighted average rule where distinct weights were assigned to individual prediction models according to their performance.

However, one of the main challenges in score fusion-like approach is finding the correct weighting for the predictions **Fig. 1** Domain generalization: The source domains are trained with diverse sets of synthetic images, and the meta-learner leverages this diversity to acquire complementary information, enabling robust generalization to unseen target distributions



of various models to minimize the impact of noise on the final prediction, as ensembles may include models that are sensitive to noise or data outliers (Shahhosseini et al., 2022). Furthermore, the main challenge in depth-based ensemble learning arises from the assumption that one of the training domains closely resembles the target domain where the model will be deployed. However, this assumption may not always hold true. Indeed, the target domain may exhibit differences in data distribution, feature representation, or other characteristics compared to the training domains. This domain shift can make it difficult for the model to generalize to previously unseen data, potentially leading to a decline in performance.

In contrast to score-based fusion or depth-based ensemble learning, Fatemifar et al. (2020) introduced a client-specific anomaly-based stacking ensemble method, where multiple deep Convolutional Neural Network (CNN) models were trained on various facial regions in an unseen attack scenario. However, implementing client-specific stacking poses a significant challenge due to the necessity of training multiple deep CNNs on various facial regions, which requires substantial amounts of labeled data. Specifically, obtaining labeled data for each facial region and for multiple clients can be laborious and may not always be attainable. Another challenge is associated with the availability of data for models that are not adequately trained or optimized for real-world deployment. This is due to the fact that real-world environments are inherently complex and dynamic, with variations that may not have been present or adequately represented

in the training data. These unforeseen variations can pose challenges for approaches that rely on ensemble learning.

#### **1.3 Limitations in Global Motion Methods**

Approaches to global motion estimation are typically categorized into two main categories: feature-based methods and direct (or featureless) methods. Feature-based methods involve detecting key points, edges, or other significant features in the data, whereas direct methods, also known as featureless methods, operate on the raw data by directly minimizing the difference between the intensities or pixel values of the images (Déniz et al., 2011). In feature-based global motion estimation, the objective is to estimate a transformation (e.g., affine or homography) that aligns the source image with the target image. However, this transformation may cause parts of the target frame to remain empty, leading to black borders around the transformed image (Muhammad et al., 2022a). This occurs because each successive transformation, such as rotations, translations, scaling, or affine changes, modifies the spatial characteristics of the image. These cumulative alterations can result in the transformed image exceeding the dimensions of the original bounding box.

The black border artifacts can be removed by first compensating for global motion using a feature-based method and then applying a frame difference approach. This combination is shown to enhance the overall motion estimation in Zhang et al. (2023). Although the dense-based feature extraction approach was employed to address the black framing issue in face PAD, it requires the extraction and processing of every pixel or grid point in the image compared to sparse methods, resulting in significantly higher computational demand (Muhammad et al., 2022b). We argue that existing approaches in face PAD often overlook the need to address this issue, which could result in the loss of critical details crucial for analyzing live and spoofed attacks. Moreover, black borders or black framing can introduce artificial cues into the video data, which might mislead or bias the models during training. Therefore, effectively managing such spatiotemporal variations and addressing black framing is essential before training any models, as it is crucial for improving the performance of face PAD.

#### **1.4 Our Contributions**

To address the aforementioned issues, we first introduce a data synthesis approach that integrates temporal and spatial transformations. Our method aims to tackle the challenge of domain generalization by employing a stacking-based ensemble learning framework based on synthetic data. To achieve this, we present a video distillation technique that blends a static image with a stabilized spatiotemporal encoded image. This blending process merges the static image's details with the blurred features of the spatiotemporal encoded image, aiding models in gaining a deeper understanding of motion blur.

We assume that data in real-world applications often contain varying degrees of blur caused by motion, low frame rates, or camera shake. On the contrary, a static image, being a single frame, may not always provide complete information. Therefore, training with temporally blurred data can make the models more resilient and better prepared for such real-world conditions. Specifically, we utilize three RNNs, namely LSTM, BILSTM, and GRU, to capture different aspects of temporal blur information, potentially enhancing the ensemble's resilience in real-world scenarios. This diversity helps the meta-learner in the stacking ensemble combine these complementary strengths effectively.

We also address the issue of black framing in face PAD, which is generated by feature-based global motion estimation methods (Muhammad et al., 2022a). We observe that existing works on face PAD do not explicitly focus on compensating for this effect, which can degrade important details to analyze in live and spoofed attacks for video-based PAD detection. This is achieved through alpha composition as a post-processing step, which involves blending the transformed image with the target frame using alpha values to seamlessly merge them. This process reduces the visibility of black framing artifacts and aids the model in learning to recognize actions in the presence of motion blur. By doing so, the proposed video distillation technique not only allows the models to focus on motion attributes, but also enables them to operate on a smaller subset of frames, thereby reducing computational overhead and improving generalization for subsequent analysis. Since our video distillation scheme follows a uniform sampling approach by dividing the original video into video clips of fixed size, it provides the flexibility to easily control the sampling rate by adjusting the segment size of the video. This is important because a higher sampling rate results in a higher temporal resolution but can introduce more noise, whereas a lower sampling rate is associated with less frequent sampling, leading to a smoother representation but potentially lower temporal detail.

Intuitively, we extend our previous approach (Muhammad et al., 2022a) in the following ways: (i) We introduce a new data augmentation technique as a post-processing step to seamlessly composite the transformed image back into the target frame, thereby reducing the visibility of black framing artifacts. These artifacts are unwanted black borders or areas that can appear around images; (ii) We address the domain generalization issue by learning from the diversity of the proposed synthetic data and introducing a deep ensemble learning framework; (iii) We use several explainability methods to answer questions such as "why did the model make a particular prediction?" or "what features were most influential in the decision-making process?"; and (iv) We balance the computational cost of the global motion estimation and system performance.

In summary, our key contributions can be summarized as follows:

- 1. We introduce a video-based data augmentation mechanism by considering both the spatial and temporal domains of the video. The proposed approach can assist deep learning models in capturing spatiotemporal information and enhancing their performance in face PAD tasks.
- 2. A meta-model is presented that leverages information from different subsets of synthetic samples, leading to improvements in the overall performance and robustness of the ensemble model.
- 3. Explainability techniques, which include gradient-weighted class activation mapping, occusion sensitivity maps and LIME, are employed to explain the decisions made by the employed model. The model reveals that motion cues are the most important factors for distinguishing whether an input image is spoofed or not.
- Experiments on seven benchmark datasets show that our proposed method provides very competitive performance in comparison with other state-of-the-art generalization methods used nowadays.

The rest of this work is organized as follows: Sect. 2 discusses recent developments and related past works,

highlighting their advantages and disadvantages. Section 3 details the various steps of our proposed method. Section 4 emphasizes the implementation details, ablation study, and comparison against several public benchmark datasets. Section 5 concludes the entire work and provides suggestions for future research.

#### 2 Literature Review

Over the past few years, face PAD methods have received considerable attention from both academia and industry. In general, these methods can be roughly categorized into appearance-based methods, temporal-based methods, and domain generalization methods.

Appearance-based methods: Conventional appearancebased techniques generally rely on extracting hand-crafted or low-level features developed prior to the emergence of deep learning. These methods involve manual design and engineering of algorithms and features, with a primary focus on analyzing static visual attributes like textures, colors, and shapes within an image or frame to make decisions. For instance, Boulkenafet et al. (2016) claimed that color information is essential for effective face presentation attack detection (PAD). They discovered that using luminancechrominance color spaces enhances detection performance compared to RGB and grayscale representations. These color spaces allowed the method to more effectively exploit the differences in color information between genuine and spoofed faces, thereby improving the overall effectiveness of the detection system. In another subsequent study (Boulkenafet et al., 2016), the authors proposed deriving a new multiscale space for image representation prior to texture feature extraction. This is achieved through the application of three multiscale filtering methods: Gaussian scale space, Difference of Gaussian scale space, and Multiscale Retinex. A robust face spoof detection algorithm using image distortion analysis (IDA) was introduced in Wen et al. (2015), considering features like specular reflection, blurriness, chromatic moment, and color diversity. Next, an ensemble of SVM classifiers, each trained for different spoof attacks (printed photos and replayed videos), was then employed to distinguish genuine faces from spoofed ones. In Yang et al. (2013), a component-based coding framework was proposed. This consists of four steps: locating face components, encoding low-level features for each component, creating a high-level face representation using Fisher criterion-based pooling, and concatenating histograms from all components to form a classifier for identification. Freitas Pereira et al. (2012) advocated the use of LBP-TOP descriptor to enhance face PAD detection. This descriptor integrates spatial and temporal information into a cohesive representation. By extending the analysis into the time domain, notable improvements over previous static frame methods were reported. Additionally, in Patel et al. (2016), image distortions were examined using various intensity channels (R, G, B, and grayscale) and across different image regions (entire image, detected face, and the facial area between the nose and chin). In Li and Feng (2019), traditional handcrafted features were combined with convolutional neural networks (CNNs) to enhance face PAD. Lately, a hybrid technique was presented in Muhammad et al. (2019), combining information on appearance from two CNNs, with an SVM classifier employed to distinguish between live and spoofed images. While appearance-focused methods exhibit enhanced performance, especially in intradatabase testing, each method has its own advantages and disadvantages. We summarize them in Table 1.

Temporal-based methods: Temporal-based methods focus on the temporal dynamics or changes occurring over a sequence of frames or in a video. These methods often involve techniques such as optical flow analysis, 3D CNN, or recurrent neural networks (RNNs) to capture temporal dependencies and patterns in the data. For instance, in Yin et al. (2016), a dense optical flow scheme was proposed to estimate the motion between successive frames. The authors reported that real and attack videos exhibit different optical flow motion patterns, enhancing the performance of face anti-spoofing detection. In Bharadwaj et al. (2013), a novel method for spoofing detection in face videos was introduced by utilizing motion magnification. It enhances facial expressions through Eulerian motion magnification and proposes two feature extraction algorithms: an improved LBP configuration for texture analysis and a motion estimation technique using the histogram of the oriented optical flow (HOOF) descriptor. To enhance the robustness of face recognition systems against spoof attacks, this paper employs dynamic mode decomposition (DMD) to capture liveness cues, such as blinking and facial dynamics (Tirunagari et al., 2015). To address 3D mask presentation attack detection (PAD), remote Photoplethysmography (rPPG) was used as an intrinsic cue, unaffected by mask material or quality. In particular, temporal variations of rPPG signals were extracted using a multi-channel time-frequency analysis scheme to enhance discriminability (Liu et al., 2021). A sample learning based recurrent neural network (SLRNN) was introduced to capture both appearance and temporal cues in Muhammad et al. (2019). In Chang et al. (2023), a Geometry-Aware Interaction Network (GAIN) using dense facial landmarks with a spatio-temporal graph convolutional network (ST-GCN) was introduced to improve the PAD detection. A generalized deep feature representation by incorporating both spatial and temporal information using a tailored 3D convolutional neural network was put forward in Li et al. (2018). The network is initially trained with augmented facial samples using crossentropy loss and further refined with a custom generalization loss serving as regularization. In Liu et al. (2018), a CNN-

 Table 1
 Summary of

 advantages and disadvantages of

appearance-based methods

International Journal of Computer Vision (2024) 132:5759–5782

Methods	Advantages	Disadvantages
Freitas Pereira et al. (2012)	By taking into account temporal dynamics, the LBP-TOP descriptor is capable of capturing motion or temporal variations	Variations in lighting conditions, camera motion, or occlusions over time could introduce inaccuracies or false positives in the analysis
Yang et al. (2013)	By pooling features with weights derived from the Fisher criterion, the method is likely to capture discriminative information	The effectiveness of feature pooling with weights derived from the Fisher criterion relies or the accurate determination of parameters
Wen et al. (2015)	Using an ensemble of SVM classifiers, enhances the algorithm's robustness and adaptability to different spoofing attacks	Extracting multiple features (specular reflection, blurriness, chromatic moment, and color diversity) can be computationally intensive
Boulkenafet et al. (2016)	The method addresses the challenge of varying image quality and resolution	Deriving a multiscale space through multiscale filtering methods may require additional computational resources
Patel et al. (2016)	By examining individual intensity channels (R, G, B, and grayscale), the method can capture subtle differences in color	Examining multiple channels and regions might lead to redundant information
Boulkenafet et al. (2016)	Color information can make the system more robust to variations in lighting and other environmental conditions	Processing and analyzing additional color space components can require more computational power
Li and Feng (2019)	Demonstrating the robustness of handcrafted features and the ability of CNNs to learn complex patterns	Integrating handcrafted features with CNNs requires careful feature engineering

RNN model was trained with pixel-wise supervision for estimating face depth and sequence-wise supervision for estimating rPPG signals, which were then fused to distinguish live versus spoof faces. In Muhammad and Oussalah (2023), a video processing scheme using a Gaussian weighting function (GWF) was proposed to model long-range temporal variations, followed by employing a CNN-RNN for PAD detection. A global motion was estimated to compensate for camera motion in Muhammad et al. (2022a), allowing for more detailed analysis of the video content by capturing subtle variations and movements. Following this, a CNN-RNN model was employed to detect PAD. This combination leverages the strengths of both global motion estimation and deep learning models. To further improve the performance of global motion, a dense sampling approach was applied in Muhammad et al. (2022b). Although temporal-focused methods typically demonstrate their effectiveness, their detection performance remains vulnerable to degradation due to realworld variations, such as user demographics, input cameras, and variations in illumination. In addition, each method possesses its own set of advantages and disadvantages, which we summarize in Table 2.

Deep Domain Generalization methods: Deep Domain Generalization (DDG) methods focus on creating models that generalize well to new unseen domains. In particular, the models aim to perform well across a variety of domains without needing to see target domain data during training. For instance, the Domain-invariant Vision Transformer (DiVT) for FAS enhanced the generalizability with respect to two loss functions. First, a concentration loss helps learn a domaininvariant representation by aggregating features of real face data. Second, a separation loss unifies various attack types from different domains (Liao et al., 2023). Huang et al. (2022) proposed adaptive vision transformers (ViT) with ensemble adapters and feature-wise transformation layers for robust cross-domain face anti-spoofing. In Liu et al. (2024), a novel Class Free Prompt Learning (CFPL) paradigm that employs two lightweight transformers, Content Q-Former (CQF) and Style Q-Former (SQF), was introduced to learn semantic prompts based on content and style features using learnable query vectors for face PAD. In Zhang et al. (2019, 2020),

 Table 2
 Summary of

 advantages and disadvantages of

temporal-based methods

Methods	Advantages	Disadvantages
Bharadwaj et al. (2013)	By using Eulerian motion magnification, subtle facial expressions are amplified, improving the accuracy of detecting spoofing attempts	The effectiveness of motion magnification and subsequent analysis relies heavily on the quality of the captured video
Tirunagari et al. (2015)	The use of dynamic mode decomposition (DMD) efficiently captures liveness cues such as blinking and facial dynamics	The accuracy of the method heavily relies on the quality of the video input
Yin et al. (2016)	Leveraging the distinct motion patterns in optical flow vectors	Calculating dense optical flow for successive frames can be computationally intensive
Muhammad et al. (2019)	The use of sparse filtering with Residual Networks (ResNet) decreases feature redundancy	The performance of the model heavily relies on the effectiveness of the sparse filtering process
Liu et al. (2021)	Using remote Photoplethysmography (rPPG) as a detection cue ensures that the method is effective regardless of the mask's material	Despite improvements, the method may still struggle with severe noise in the rPPG signals, affecting its accuracy
Muhammad et al. (2022a)	Compensates for camera movement, providing a more stable and smoother video sequence for subsequent analysis	The appearance of black framing and other artifacts can introduce biases into the training data
Chang et al. (2023)	The cross-attention feature interaction mechanism allows GAIN to integrate seamlessly with other existing methods	The use of dense facial landmarks, ST-GCN, and cross-attention mechanisms can make the model complex to implement
Muhammad et al. (2022b)	Dense sampling can help avoid black framing after transformations	Significantly raises the computational demands, requiring more processing power

a multi-modal fusion approach, aiming to enhance generalization by conducting feature re-weighting, emphasizing the most informative channel features while attenuating the less relevant ones within each modality, was investigated. To extract discriminative fused features, a Partially Shared Multi-modal Network is proposed to learn the fused information from single-modal and multi-modal branches (Liu et al., 2021). For the same purpose, George et al. suggested that incorporating analysis across multiple channels could offer a solution to this problem. Hence, they introduced a CNN approach that operates across multiple channels for PAD (George et al., 2019). In Srivatsan et al. (2023), a new method for making FAS more robust across different situations was introduced by connecting visual representations with natural language. By matching image features with descriptions, based on how we talk about things, FAS can work better even when there's not much data available. In Wang et al. (2020), it was found that incorporating spatio-temporal information from multiple frames using a Spatio-Temporal Propagation Module (STPM) can help the model generalize better to variations in depth cues

across different environments or conditions. An Instance-Aware Domain Generalization framework was introduced in Zhou et al. (2023) by aligning features at the instance level without domain labels. Sun et al. (2023) proposed Dynamic Kernel Generator and Categorical Style Assembly to extract instance-specific features for improving generalization of face PAD. Specifically, the authors formulate their FAS strategy of separability and alignment (SA-FAS) as a problem of invariant risk minimization (IRM), and proposed encouraging domain separability while ensuring uniform live-to-spoof transitions across domains. Inspired by vision-language models, a method known as VL-FAS was introduced in Fang et al. (2024) by leveraging fine-grained natural language descriptions to guide attention towards the face region, resulting in cleaner and more generalized features.

and memory

A domain adaptation method that generates pseudolabeled samples named cyclically disentangled feature translation network (CDFTN) was proposed in Yue et al. (2023). Chuang et al proposed to improve the generalization based on one-side triplet loss (Chuang et al., 2023). In Cai et al. (2022),

International Journal of Computer Vision (2024) 132:5759–5782

Table 3Summary ofadvantages and disadvantages of	Methods	Advantages	Disadvantages
domain generalization methods	Zhang et al. (2019)	The multi-modal fusion approach improves generalization by effectively combining information from multiple modalities	Despite improvements, effective multi-modal fusion often requires sufficient labeled data from each modality for training
	Wang et al. (2020)	Introducing a novel Contrastive Depth Loss can lead to more accurate depth supervision, which can be beneficial for improving generalization	The method involves multiple modules (RSGB, STPM, Contrastive Depth Loss) which may increase the complexity of the model
	Huang et al. (2022)	The ensemble adapters module enables the model to adapt to different domains with only a few samples	There is still a risk of overfitting, especially if the samples are not sufficiently representative of the target domain's variability
	Liao et al. (2023)	The combination of concentration loss and separation loss improves the model's ability to learn domain-invariant representations	Transformers typically require large amounts of data to train effectively
	Srivatsan et al. (2023)	Leveraging both images and text through multimodal contrastive learning enhances feature generalization	There might be a semantic gap between the natural language descriptions and the actual image content
	Zhou et al. (2023)	Asymmetric Instance Adaptive Whitening, Dynamic Kernel Generator, and Categorical Style Assembly offer flexibility and adaptability in handling various styles and shifts in the data	Optimizing the various components and hyperparameters of the framework may require careful tuning and additional complexity
	Liu et al. (2024)	Using lightweight transformers, CQF and SQF, makes the model efficient	Integrating and fine-tuning multiple components such as CQF, SQF, PTM, and DSP adds complexity to the model's implementation

a two-stream network was utilized to fuse the input RGB image, while meta-pattern learning was proposed to improve generalization. In Huang et al. (2022), a cross-adversarial training scheme was proposed to improve the generalization by minimizing the correlation among two sets of features. The work reported in Zhou et al. (2022) aims to learn a generalized feature space by designing the target data to the source-domain style called generative domain adaptation (GDA). A hypothesis verification framework was proposed in Liu et al. (2022) where two hypothesis verification modules are utilized for improving the generalization. In Wang et al. (2022), a novel shuffled style assembly network (SSAN) was introduced by aligning multiple source domains into a stylized feature space, while domain generalization was improved by a contrastive learning strategy. To select a common feature space, adversarial learning was proposed, and aggregation of live faces was performed to achieve a generalized feature space in Liu et al. (2022). Nonetheless, there is no clear consensus that the pre-defined distributions can be considered the optimal ones for the feature space. We argue that different domains have their strengths and weaknesses. By combining them, their collective knowledge and diversity can lead to a more comprehensive understanding of faces. However, implementing and fine-tuning such approaches, as discussed in Table 3, might be complex and require careful consideration of alignment methods, ensemble learning techniques, and evaluation strategies to ensure the desired performance improvements are achieved.

### **3 Methodology**

A key idea underlying our approach is to enhance the PAD model's generalization ability. This objective is achieved through training multiple sub-models using distinct subsets of synthetic data. Specifically, we embark on data augmentation by blending a static image with spatiotemporal encoded images, effectively replicating the visual effect of motion blur through the manipulation of alpha values applied to the static image. This variation in alpha values governs the extent of



Fig. 2 Schematic diagram of the proposed data augmentation and deep ensemble learning for face anti-spoofing countermeasure

each image's influence on the final composite, thereby yielding an array of synthetic data subsets. As a further stride in our methodology, we introduce a meta-model, which capitalizes on the predictions of the base models. In the following sections, we describe all these steps in detail. An overall view of our proposed methodology is presented in Fig. 2.

#### 3.1 Data Augmentation

The process of data augmentation involves five main steps, as follows: (1) Select a source and a target frame of an input video; (2) Estimate the global motion transformation between the source and target frames; (3) Perform the geometric transformation on the source frame to align it with the target frame; (4) Create a new spatiotemporal encoded image by stabilizing the segments of the video; and (5) Use alpha blending to combine the transformed image with the target (static) frame, considering the transparency of each pixel. Each of these steps is explained in the following subsections.

#### 3.1.1 Global Motion Estimation

Suppose a video V is equally divided into P non-overlapping segments, i.e.,  $V = \{S_k\}_{k=1}^{P}$ , where  $S_k$  is the k-th segment. The length of each segment is set to be N = 40 frames. For each segment, features are extracted from the fixed (first) and moving (second) images of the segment. In particular, the FAST feature detector (Rosten & Drummond, 2005) is utilized to detect interest points, and then the FREAK descriptor (Alahi et al., 2012) extracts the features to collect points of interest from both frames. For matching the interest points, Hamming distance (HD) is utilized in our work. The transformation between frames is calculated from the first frame onward using a rigid (Euclidean space) transformation. The rigid transformation preserves lengths, angles, and shapes, and includes translation, rotation, and reflection. On the contrary, the affine transformation used in Muhammad et al. (2022a) includes translation, rotation, scaling, and shearing, and preserves parallel lines and ratios of distances, but can change shapes through scaling and shearing. Thus, rigid transformation for each subsequent frame in the segment is

estimated as:

$$\begin{bmatrix} a'\\b'\\1 \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & d_{13}\\d_{21} & d_{22} & d_{23}\\d_{31} & d_{32} & d_{33} \end{bmatrix} \begin{bmatrix} a\\b\\1 \end{bmatrix} = M \begin{bmatrix} a\\b\\1 \end{bmatrix}$$
(1)

In this equation, [a' b' 1] represents the homogeneous coordinates in the fixed image, and [a b 1] represents the homogeneous coordinates in the moving image. The rigid transformation matrix M describes the inter-frame motion. To eliminate false-matching points and robust estimation of the geometric transformation between the frames, we use the M-estimator Sample Consensus (MSAC) algorithm (Torr & Zisserman, 1997) to detect outliers and remove false matching points. To obtain a warped (spatiotemporal) image, accumulation is performed using the following equation (Muhammad et al., 2022b):

$$A = \frac{1}{N} \sum_{i=1}^{N} I_i \cdot M, \tag{2}$$

where *A* is a single image that represents the combined motion in the segment after applying the accumulated transformation *M*, *N* is the number of frames that are considered for accumulation and  $I_i$  represents the *i*-th frame in the segment. Each frame  $I_i$  is transformed using the accumulated transformation *M* before being added to the sum.  $\frac{1}{N}$  is the scaling factor that ensures the final result is an average. In particular, Eq. (2) shows that we take each frame in the segment, apply the accumulated transformation *M* to it, sum up these transformed frames, and then divide by *N* to generate a final composite frame *A* for that particular segment. To show the importance of this approach, we simply remove the accumulated transformation *M* and use alternatively the following equation:

$$B = \frac{1}{N} \sum_{i=1}^{N} I_i.$$
 (3)

This calculates the average frame B by summing all the individual frames in the segment and then dividing that sum by the total number of frames N. The outcome of both equations is illustrated in Fig. 3. The first row represents the noisy camera motion, which can be observed after using the Eq. (3). The second row shows that the proposed approach significantly removes the motion blur using the Eq. (2). Although the camera motion issue is eliminated, black framing near the border of the images in the second row appears, which requires further preprocessing. In the following sub-section, we address this issue and explain the motivation of data augmentation.



**Fig.3** a Temporal averaging is performed based on Eq. (3) to visualize the global motion, which exhibits a significant amount of distortion in the encoded image. b The cumulative transformation is calculated by applying the transformation matrix of each frame's motion to the previous cumulative transformation, according to Eq. (2). c The result of the cumulative transformation with alpha blending is based on Eq. (4). The images in the column are the results of averaging 40 frames

#### 3.1.2 Alpha Transparency

While Eq. (2) effectively removes the disruptive effects of noisy camera motion, the spatiotemporal-encoded images inherently exhibit a certain level of motion blur. This blur arises from the inherent movement between the camera and the scene during each exposure. In particular, if the subjects move their hands or change the direction of their face in the scene while the camera's shutter is open, their motion appears as a blur in the spatiotemporal-encoded image. This observation has encouraged us to propose a new basic augmentation technique, called alpha transparency, for face PAD. The concept is to assign appropriate alpha values to the first (static) frame of the segment and then blend it with the spatiotemporal-encoded image based on those alpha values. This approach helps preserve more details and finer features of the scene, as the spatiotemporal-encoded image captures different stages of the motion blur. The alpha blending process is achieved through the following two steps: (1) Obtain a source image (i.e., a spatiotemporal-encoded image based on Eq. (2)); and (2) Choose first (still) image of each segment to blend with the source image. Let us assume, without loss of generality, that we blend the source image  $P_1$  over the target image  $P_2$  as follows:

$$P_{blended}(a,b) = \alpha P_1(a,b) + (1-\alpha)P_2(a,b).$$
(4)

The equation represents how to combine the information from the two source values  $P_1(a, b)$  and  $P_2(a, b)$  to create a new value at the location (a, b) in the synthetic image. The weight factor  $\alpha$  determines the contribution of  $P_1(a, b)$  to the blended result. When  $\alpha$  is closer to 1,  $P_1(a, b)$  has a higher influence, and when  $\alpha$  is closer to 0,  $P_2(a, b)$  has a higher influence. Thus, blending the source image with the target image helps to eliminate the black framing issue, as shown in the third row of Fig. 3, creating a visually seamless transition. Moreover, Fig. 4 displays images that represent the results of blending with different  $\alpha$  values. This approach can be useful in scenarios where someone wants to emphasize the overall motion of the scene while maintaining a recognizable background or main subject. We hypothesize that the combination of alpha blending and spatiotemporal encoding can be beneficial in simulating motion blur and helping the deep learning model become more robust to real-world scenarios where camera motion is present.

#### 3.2 Deep Ensemble Learning

Deep learning methods based on 2D Convolutional Neural Networks (CNNs) have shown improved performance compared to classical machine learning approaches (Wang et al., 2020; Liu et al., 2021; Shao et al., 2020). However, mainstream 2D CNN frameworks primarily focus on spatial information, thus lacking the capacity to understand sequential data. Moreover, 2D CNNs do not possess a memory mechanism to capture temporal relations. Motivated by the fact that recurrent neural networks (RNNs) can effectively handle temporal information, we develop a stacking-based deep ensemble learning framework to learn from motion blur by processing sequences of images representing motion over time. Ensemble learning has been supported by multiple approaches such as bagging, boosting, and stacking, resulting in a better generalization of learning models (Fatemifar et al., 2020). Specifically, stacking is an integration technique that involves combining the predictions based on different weak model predictions, where the meta-model is used to integrate the output of base models (Ganaie et al., 2021). One common approach in stacked ensemble learning is to develop a set of Tier-1 classifiers denoted as  $S_1, S_2, S_3, \ldots, S_N$ . These classifiers are developed through cross-validation of the training dataset. Then, the outputs or predictions from these Tier-1 classifiers are subsequently employed as inputs for a higherlevel classifier or meta-learner within the ensemble (Polikar, 2012).

Since our primary goal is to learn from the variations of motion blur sequences, we train diverse recurrent neural networks to leverage the benefits of the proposed data augmentation mechanism. The approach involves fine-tuning a CNN model on the labeled datasets in the first stage. Then, we extract the fine-tuned features from the pooling layer and utilize them as input to train different variants of RNNs. In our work, three base models-Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), Bidirectional Long Short-Term Memory (BiLSTM) (Schuster & Paliwal, 1997), and Gated Recurrent Unit (GRU) (Cho et al., 2014) are selected. The first base model (LSTM) captures temporal patterns and dynamics across frames using its special memory cells, allowing it to maintain long-range dependencies in sequential data. When presented with a sequence of blurred frames, the LSTM can learn to recognize patterns one by one and retain relationships between consecutive frames. The second base model (BiLSTM) captures information from both past and future contexts in sequential data. It consists of two sets of LSTM cells, where both cells work independently to capture information from both past and future contexts. The hidden states from both directions are then concatenated or combined to obtain the final output. The third base model (GRU) incorporates gating mechanisms similar to LSTM, enabling it to selectively learn and update information over time.

In particular, this approach encourages each submodel to focus on different aspects of the data, such as variations in temporal blur. For example, one model might specialize in certain features based on synthetic data, while another model may perform better in other aspects. We then combine the predictions from these weak experts (base models) and use them as input to a meta-model (another RNN). The metamodel learns to integrate these predictions and make the final decision. We call it a meta-model because it leverages the diversity and complementary strengths of the individual base models, leading to improved generalization for face PAD. Table 4 illustrates the configuration of the proposed metamodel.

#### 4 Experimental Analysis of Using Open Datasets

Since spoofing attacks have become more realistic and a bigger security concern for face recognition systems, researchers around the world have paid attention to developing more diverse datasets. These datasets include faces captured in various lighting conditions and utilize different spoofing techniques, such as photos, videos, makeup, masks, etc. These diverse datasets can be single modality, focusing on one type of data (e.g., images), or multimodal, combining multiple data types (e.g., images and depth information). To evaluate our model's performance in detecting face spoofing attacks, we trained and tested it on several state-of-the-art datasets widely used in the face anti-spoofing domain. Details about these datasets are provided in the following section.

#### 4.1 Datasets

Seven diverse databases, including the Wide Multi-Channel Presentation Attack (WMCA) database (George et al., 2019), Fig. 4 2D synthetic samples from CASIA-MFSD are shown. In the left column, we present video segments used in the process of data augmentation. In the right column, we display composite images after blending with spatiotemporal images using alpha values of 0.5 (Synt 1), 1.0 (Synt 2), and 1.5 (Synt 3), respectively. The encoded clip is based on alpha values of 0.1. These synthetic samples can be effectively employed for ensemble stacking, resulting in a significant improvement in face anti-spoofing performance

Table 4Base modelarchitectures and their

parameters

Video sequence	Encoded clip	Synt 1	Synt 2	Synt 3

	Base model 1	Base model 2	Base model 3	Meta-Model
Layer type	LSTM	BiLSTM	GRU	BiLSTM
No. of layers	1	1	1	1
No. of units	500	500	500	100
Optimizer	ADAM	ADAM	ADAM	ADAM
Learning rate	0.0001	0.0001	0.0001	1
Cost function	Cross entropy	Cross entropy	Cross entropy	Cross entropy

the CASIA-SURF database (Zhang et al., 2020), the OULU-NPU database (denoted as O) (Boulkenafet et al., 2017), the CASIA-MFSD database (denoted as C) (Zhang et al., 2012), the Idiap Replay-Attack database (denoted as I) (Chingoyska

the Idiap Replay-Attack database (denoted as I) (Chingovska et al., 2012), the MSU-MFSD database (denoted as M) (Wen et al., 2015), and the SiW-Mv2 dataset (Guo et al., 2022) were used to conduct the experiments.

The Wide Multi Channel Presentation Attack (WMCA) database comprises 1941 short video recordings, featuring both genuine and presentation attacks from 72 distinct identities. These recordings utilize various channels, including color, depth, infrared, and thermal.

The Chalearn CASIA-SURF dataset is one of the largest face anti-spoofing datasets, consisting of 1000 subjects and 21,000 video clips across three modalities: RGB, Depth, and IR. Each sample in the dataset comprises one live video clip and six spoof video clips, each representing a different attack method.

The OULU-NPU database contains 4950 videos, encompassing both real and attack videos. Two primary presentation attack types were considered in this database: print and videoreplay. The videos recorded from the 55 subjects. CASIA-MFSD consists of a total of 50 subjects, with each subject having 12 videos captured under varying resolutions and lighting conditions. This dataset is developed to include three distinct types of spoof attacks: replay, warp print, and cut print attacks. In particular, CASIA-MFSD comprises 600 video recordings. Out of these, 240 videos from 20 subjects are allocated for training purposes, while the remaining 360 videos from 30 subjects are designated for testing.

The Idiap Replay-Attack database comprises 1300 video clips that consist of photo and video attacks on 50 clients, all conducted under varying lighting conditions. Data for the attacks was gathered in two distinct lighting conditions: controlled, with office lights on, blinds down, and a uniform background; and adverse, with raised blinds, a complex background, and no office lighting.

The MSU-MFSD dataset was constructed through the participants of 35 individuals, resulting in a combined total of 280 video entries. The recordings were captured using two distinct camera types, each with varying resolutions (720  $\times$  480 and 640  $\times$  480). Regarding generating live recordings, every participant contributed two video clips, one recorded with a laptop camera and the other with an Android device. In contrast, for video attack instances, two different camera models, specifically iPhone and Canon cameras, were utilized to capture high-definition videos for each subject. Thus, the presence of diverse lighting conditions and a wide array of attack types introduces significant complexity and difficulty when dealing with these datasets.

The SiW-Mv2 dataset includes 785 videos from 493 subjects and 915 spoof videos from 600 subjects, ranging from common print and replay attacks to various masks, impersonation makeup, and physical material coverings. It is a large-scale face anti-spoofing (FAS) dataset with 14 types of diverse spoofing attacks. For instance, obfuscation makeup and partial coverings are designed to hide a subject's identity, while impersonation makeup and masks are used to imitate other identities.

#### 4.2 Implementation Details

All the images are adjusted to a size of  $224 \times 224$  to align with the input specifications of the pretrained DenseNet-201 architecture (Huang et al., 2017). Fine-tuning of the CNN model is carried out employing the Stochastic Gradient Descent (SGD) optimizer, with a mini-batch size of 32 and a validation check performed every 30 iterations. The learning rate is set at 0.0001, and fixed-size epochs are not used. Instead, we implement an early stopping mechanism (Prechelt, 1998) to automatically stop the training process to prevent overfitting. During the ensemble learning phase, the CNN model undergoes fine-tuning with spatiotemporalencoded video clips with alpha values of 0.1 and three distinct synthetic sets individually. These subsets introduce random expansion to the training images through alpha values of 0.5, 1.0, and 1.5 as defined in Eq. (4). Additional data augmentation techniques are utilized, such as rotation within the range of -20 to +20 degrees, as well as random translations along the x and y axes. Subsequently, the features from the finetuned model are extracted and used as inputs to train a LSTM, a BiLSTM and a GRU.

#### 4.3 Evaluation Metrics

Various standard metrics are utilized in this study for comparison, such as Equal Error Rate (EER), Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER). The EER is often used in biometric systems to represent the point where the false acceptance rate (FAR) and the false rejection rate (FRR) are equal. BPCER measures the rate at which genuine presentations are incorrectly classified as attacks, while APCER measures the rate at which attack presentations (fake attempts) are incorrectly classified as genuine. ACER is the average of APCER and BPCER, providing a balanced view of the system's performance in detecting both attacks and genuine attempts. Additionally, Half Total Error Rate (HTER) is reported, which is the average of FAR and FRR after setting a specific threshold based on the validation set. The Area Under the Curve (AUC) is also reported, measuring the model's overall ability to distinguish between classes, with the area under the ROC curve representing performance across all thresholds.

#### 4.4 Comparison Against the State-of-the-Art Methods

To assess the effectiveness of the proposed method, we conducted comparisons with some of the most representative state-of-the-art methods. Tables 5 and 6 present the performance evaluation on WMCA and CASIA-SURF datasets, for intra-dataset testing scenarios. The bold values in the tables indicate the best reported results.

Intra-Testing Results: Table 5 presents a comparison of the performance of our proposed method on the WMCA dataset with other CNN-based approaches (George & Marcel, 2019; Liu et al., 2021; He et al., 2016; Tolstikhin et al., 2021; Wang et al., 2022) and Transformer-based method (Antil & Dhiman, 2024). Similar to the previous approach described in Antil and Dhiman (2024), our proposed metamodel achieves the best performance with the lowest ACER (0.1%)in the intra-testing scenario. Furthermore, the metamodel has a slightly better BPCER (0.0%) compared to the T-Encoder, which indicates perfect recognition of bona fide presentations. Table 6 provides the performance comparison on the CASIA-SURF dataset. The comparison methods (Liu et al., 2021; Zhang et al., 2019, 2020; Yu et al., 2023) employ multimodal approaches. Among these, the proposed meta-model stands out for achieving the lowest ACER of 0.5%. The experimental findings validate that our performance is consistent with other leading benchmarks. Thus, it shows that stacking-based ensemble learning is well-suited for intra-testin scenario.

*Cross-Dataset Testing:* Since our work focuses on improving the generalization of face PAD, we initially conducted cross-dataset testing experiments between two datasets, namely, the WMCA and CASIA-SURF datasets. In particular, we begin by training the model using the CASIA-SURF dataset and then assess its performance on the WMCA dataset, focusing on HTER. Similarly, we invert the process by training the model on the WMCA dataset and evaluating it on the CASIA-SURF dataset. Table 7 showcases the comparison of performances, highlighting the meta-model's attainment of state-of-the-art results on the WMCA dataset. In particular, when trained on CASIA-SURF and tested on WMCA, the meta-model shows an error rate of 17.76%, indicating its robustness in cross-database evaluation. When trained on WMCA and tested on CASIA-SURF, ViT+AMA

Table 5         Performance           comparison on the WMCA	Methods	APCER	BPCER	ACER
dataset during intra-dataset	DeepPixBis (George & Marcel, 2019)	8.2	3.7	6.0
testing (%)	MA-Net (Liu et al., 2021)	11.1	2.6	6.8
	ResNet (He et al., 2016)	3.5	1.6	2.6
	LBP-SVM (Chingovska et al., 2012)	8.5	0.6	4.6
	MLP-Mixer (Tolstikhin et al., 2021)	1.7	2.3	2.0
	Conv-MLP (Wang et al., 2022)	0.8	1.0	<u>0.9</u>
	T-Encoder (Antil & Dhiman, 2024)	0.1	<u>0.1</u>	0.1
	Meta-Model (ours)	<u>0.2</u>	0.0	0.1
	Best results appear in bold, while the second-	best are marked with an	underline	

The results are obtained directly from the cited paper(Antil & Dhiman, 2024)

 
 Table 6
 Performance
 comparison on the CASIA-SURF dataset during intra-dataset testing (%)

Table 7 Ablation study using

cross-database evaluation

Methods	APCER	BPCER	ACER
Halfway Fusion (Zhang et al., 2019)	5.6	3.8	4.7
SE Fusion (Zhang et al., 2019)	3.8	1.0	2.4
Zhang et al. (2020)	2.8	0.3	1.5
MA-Net (Liu et al., 2021)	2.4	1.7	2.0
Conv-MLP (Wang et al., 2022)	1.5	1.8	1.6
ViT+AMA (Yu et al., 2023)	<u>0.8</u>	0.4	<u>0.6</u>
T-Encoder (Antil & Dhiman, 2024)	1.6	1.2	1.4
Meta-Model (ours)	0.4	0.7	0.5

Best results appear in bold, while the second-best are marked with an underline The results are obtained directly from the cited paper (Antil & Dhiman, 2024)

#### Method Train: CASIA-SURF Train: WMCA Test: WMCA Test: CASIA-SURF 24.54 Aux. (Depth) (Liu et al., 2018) 12.35 MM-CDCN (Yu et al., 2020) 21.83 21.25 MA-ViT (Liu et al., 2021) 20.63 10.41 ViT (Dosovitskiy et al., 2020) 23.21 19.19 ViT+AMA (Yu et al., 2023) 18.83 8.60 T-Encoder (Antil & Dhiman, 2024) 19.43 13.84 17.76 14.20 Meta-Model (ours)

Best results appear in bold, while the second-best are marked with an underline

The results are obtained directly from the cited paper (Antil & Dhiman, 2024)

(Yu et al., 2023) achieves the lowest error rate of 8.60%, making it the best performer in this scenario.

We extend our experiments to the commonly employed cross-dataset testing scenario, where the model is trained on three source databases and evaluated on a completely unseen database using the leave-one-out (LOO) strategy. Specifically, four datasets - the OULU-NPU database (denoted as O), CASIA-MFSD database (denoted as C), Idiap Replay-Attack database (denoted as I), and MSU-MFSD database (denoted as M)-are employed in various combinations: O&C&I to M, O&M&I to C, O&C&M to I, and I&C&M to O. Table 8 presents a performance comparison with recently proposed state-of-the-art methods in terms of HTER and AUC. It can be observed that the meta-model provides the best results for three protocols: O&C&I to M, O&M&I to C, and I&C&M to O. For O&C&I to M, the meta-model has the lowest HTER (1.20). Similarly, the meta-model has the lowest HTER (1.37) for O&M&I to C. In the case of I&C&M to O, the meta-model has the lowest HTER (2.08). However, DiVT-M performs best with the lowest HTER (3.71), outperforming the meta-model in O&C&M to I. Overall, our proposed method demonstrates itself as the most effective method across the majority of the evaluated scenarios, showcasing its robustness and superior performance in crossdatabase face presentation attack detection.

Table 8	Performance evaluation	using MSU-MFSD	(M), Idiap (I),	CASIA (C),	and OULU-NPU	(O) databases
---------	------------------------	----------------	-----------------	------------	--------------	---------------

	O&C&I t	o M	O&M&I	to C	O&C&M	to I	I&C&M	to O
Methods	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
MADDG (Shao et al., 2019)	17.69	88.06	24.50	84.51	22.19	84.99	27.89	80.02
DAFL (Saha et al., 2020)	14.58	92.58	17.41	90.12	15.13	95.76	14.72	93.08
SSDG-R (Jia et al., 2020)	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54
DR-MD (Wang et al., 2020)	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47
RFMetaFAS (Shao et al., 2020)	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
FAS-DR-BC(MT) (Qin et al., 2021)	11.67	93.09	18.44	89.67	11.93	94.95	16.23	91.18
ADL (Liu et al., 2022)	5.00	97.58	10.00	96.85	12.07	94.68	13.45	94.43
HFN + MP (Cai et al., 2022)	5.24	97.28	9.11	96.09	15.35	90.67	12.40	94.26
Cross-ADD (Huang et al., 2022)	11.64	95.27	17.51	89.98	15.08	91.92	14.27	93.04
GDA (Zhou et al., 2022)	9.20	98.00	12.20	93.00	10.00	96.00	14.40	92.60
Regression Network (Kwak et al., 2023)	5.41	98.85	10.05	94.27	8.62	97.60	11.42	95.52
SSAN-R (Wang et al., 2022)	6.67	98.75	10.00	96.67	8.88	96.79	13.72	93.63
FG + HV (Liu et al., 2022)	9.17	96.92	12.47	93.47	16.29	90.11	13.58	93.55
CRFAS (Zheng et al., 2023)	7.14	97.44	9.88	96.56	8.57	96.07	16.38	90.87
DiVT-M (Liao et al., 2023)	2.86	99.14	8.67	96.92	<u>3.71</u>	99.29	13.06	94.04
ViTAF (Huang et al., 2022)	2.92	<u>99.62</u>	1.40	99.92	1.64	<u>99.64</u>	5.39	98.67
FLIP-MCL (Srivatsan et al., 2023)	4.95	98.11	0.54	<u>99.98</u>	4.25	99.07	<u>2.31</u>	<u>99.63</u>
IADG (Zhou et al., 2023)	5.41	98.19	8.70	96.44	10.62	94.50	8.86	97.14
SA-FAS (Sun et al., 2023)	5.95	96.55	8.78	95.37	6.58	97.54	10.00	96.23
VL-FAS (Fang et al., 2024)	3.13	99.31	4.00	98.64	5.00	98.90	7.92	97.05
Meta-Model (ours)	1.20	<b>99.98</b>	<u>1.37</u>	99.99	5.40	<b>99.97</b>	2.08	<b>99.77</b>

Best results appear in bold, while the second-best are marked with an underline

The results are obtained directly from the cited papers

We also provide a more comprehensive evaluation of the classifier's performance using the Area Under the ROC Curve (AUC) shown in Table 8. Although the meta-model provides low performance on one database (i.e., Replay-Attack), one can see that the meta-model achieves more than 99% AUC on all the datasets. Since EER is calculated based on the testing set of source databases, HTER focuses on finding the operating point (threshold) where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal. If the AUC is high and the HTER is low, it indicates that the classifier achieves a good balance between FAR and FRR. Since AUC is not threshold-dependent, it is more useful for comparing different classifiers or evaluating the model's generalization capability.

To quantitatively assess various types of spoofing attacks, we implement 13 leave-one-out testing protocols on the SiW-M dataset. According to Liu et al. (2019), the model is trained using 12 types of spoof attacks along with 80% of the live videos, and evaluated on the remaining attack type and the other 20% of live videos. The results are analyzed in Table 9 with two of the most recent face anti-spoofing methods (Liu et al., 2019, 2020), using (Liu et al., 2018) as the baseline, as it has demonstrated state-of-the-art performance across various

benchmarks. In comparison with other methods, the metamodel performs best on several attacks, including Replay Attack, Print Attack, Paper 3D Mask, Paper Glasses Partial Attack, and Partial Fun Eye Glasses Attack. Specifically, it achieves the lowest APCER (0.1), BPCER (7.8), ACER (3.9), and EER (1.1) for Replay Attack. For Print Attack, it consistently has the lowest BPCER (2.5), ACER (1.2), and EER (0.6). In the case of Paper 3D Mask, it has the lowest APCER (0.0) and ACER (3.2), along with the second-lowest BPCER (6.4). Similarly, it records the lowest BPCER (0.0), ACER (0.0), and EER (0.0) for Paper Glasses Partial Attack, and the lowest BPCER (0.0), ACER (14.8), and EER (10.9) for Fun Partial Attack. Thus, the meta-model demonstrates superior performance, particularly excelling in Print and Paper Glasses attacks. It achieves the lowest average BPCER (6.6  $\pm$  5.9), ACER (13.7  $\pm$  10.8), and EER (9.3  $\pm$  7.4), indicating its robustness across various attack types.

#### 4.5 Experiment on Limited Source Domains

Our study also investigates the scenario of a limited source domain by training the model on only two source domains, as opposed to the three domains used in Table 8. It can be

			3D Mas	k				Makeup			Partial a	ttacks		
Methods	Replay	Print	Half	Silic.	Trans.	Paper	Manne.	Ob.	Im.	Cos.	Fun.	Papergls.	Paper	Average
APCER														
LBP+SVM (Boulkenafet et al., 2017)	19.1	15.4	40.8	20.3	70.3	0.0	4.6	96.9	35.3	11.3	53.3	58.5	0.6	$32.8\pm29.8$
Auxiliary (Liu et al., 2018)	23.7	7.3	27.7	18.2	97.8	8.3	16.2	100.0	18.0	16.3	91.8	72.2	0.4	$38.3 \pm 37.4$
DTL (Liu et al., 2019)	1.0	0.0	0.7	24.5	58.6	0.5	3.8	73.2	13.2	12.4	17.0	17.0	0.2	$17.1 \pm 23.3$
DSF (Liu et al., 2020)	1.6	0.0	0.5	7.2	9.7	0.5	0.0	96.1	0.0	21.8	14.4	6.5	0.0	$12.2\pm 26.1$
Meta-Model (ours)	0.1	0.0	20.4	21.6	33.4	0.0	11.5	79.2	21.7	22.5	29.7	30.5	0.0	$20.8\pm20.6$
BPCER														
LBP+SVM (Boulkenafet et al., 2017)	22.1	21.5	21.9	21.4	20.7	23.1	22.9	21.7	12.5	22.2	18.4	20.0	22.9	$21.0 \pm 2.9$
Auxiliary (Liu et al., 2018)	10.1	6.5	10.9	<u>11.6</u>	6.2	7.8	<u>9.3</u>	11.6	9.3	7.1	<u>6.2</u>	8.8	10.3	$\underline{8.9} \pm 2.0$
DTL (Liu et al., 2019)	18.6	11.9	29.3	12.8	13.4	8.5	23.0	11.5	9.6	16.0	21.5	22.6	16.8	$16.6\pm 6.2$
DSF (Liu et al., 2020)	14.0	14.6	13.6	18.6	18.1	<u>8.1</u>	13.4	10.3	9.2	17.2	27.0	35.5	11.2	$16.2\pm7.6$
Meta-Model (ours)	7.8	2.5	3.1	5.6	16.0	6.4	7.6	11.4	4.2	1.6	0.0	0.0	20.6	$6.6\pm5.9$
ACER														
LBP+SVM (Boulkenafet et al., 2017)	20.6	18.4	31.3	21.4	45.5	11.6	13.8	59.3	23.9	16.7	35.9	39.2	11.7	$26.9\pm14.5$
Auxiliary (Liu et al., 2018)	16.8	6.9	19.3	14.9	52.1	8.0	12.8	55.8	13.7	11.7	49.0	40.5	5.3	$23.6\pm18.5$
DTL (Liu et al., 2019)	9.8	6.0	15.0	18.7	36.0	4.5	13.4	48.1	11.4	14.2	19.3	19.8	8.5	$16.8\pm11.1$
DSF (Liu et al., 2020)	7.8	7.3	7.1	12.9	13.9	4.3	6.7	53.2	4.6	19.5	20.7	21.0	5.6	$15.0 \pm 12.0$
Meta-Model (ours)	3.9	1.2	11.7	<u>13.6</u>	24.7	3.2	<u>9.5</u>	45.3	12.9	12.1	14.8	15.2	10.3	$13.7 \pm 10.8$
EER														
LBP+SVM (Boulkenafet et al., 2017)	20.8	18.6	36.3	21.4	37.2	7.5	14.1	51.2	19.8	16.1	34.4	33.0	7.9	$24.5\pm12.9$
Auxiliary (Liu et al., 2018)	14.0	4.3	11.6	12.4	24.6	7.8	10.0	72.3	10.1	9.4	21.4	18.6	4.0	$17.0 \pm 17.7$
DTL (Liu et al., 2019)	10.0	2.1	14.4	18.6	26.5	5.7	9.6	50.2	10.1	13.2	<u>19.8</u>	20.5	8.8	$16.1\pm12.2$
DSF (Liu et al., 2020)	<u>7.6</u>	3.8	8.4	13.8	14.5	5.3	4.4	35.4	0.0	19.3	21.0	20.8	<u>1.6</u>	$\underline{12.0}\pm10.0$
Meta-Model (ours)	1.1	0.6	8.7	7.0	20.8	3.1	9.4	28.9	9.3	5.3	10.9	9.8	6.5	$9.3\pm7.4$
Best results appear in bold, while the sec The results are obtained using leave-one-	cond-best ar- -out testing	e marked v (%) direct	with an un ly from th	derline e cited pa	per(Liu et a	1., 2020)								

 $\underline{\textcircled{O}}$  Springer

Table 10         The results of cross-dataset testing on limited source doma	ins
--	-----

Methods	O&I to I	M	M&I to	С	O&I to 0	2	O&M to I		C&M to	0
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
Supervised (Wang et al., 2021)	12.1	94.2	30.4	77.0	18.0	90.1	16.8	93.8	17.9	89.5
Mean-Teacher (Tarvainen & Valpola, 2017)	19.6	86.5	31.1	76.6	23.7	84.9	18.4	86.0	23.5	84.9
USDAN (Jia et al., 2021)	15.8	88.1	35.6	69.0	33.3	72.7	19.8	87.9	20.2	88.3
EPCR-labeled (Wang et al., 2021)	12.5	<u>95.3</u>	18.9	89.7	18.9	89.7	14.0	92.4	17.9	90.9
EPCR-unlabeled (Wang et al., 2021)	<u>10.4</u>	94.5	25.4	83.8	16.7	<u>91.4</u>	12.4	<u>94.3</u>	17.8	91.3
CNN-LSTM	15.2	89.5	27.8	81.3	18.5	86.2	19.2	86.7	19.8	85.3
CNN-BiLSTM	14.8	88.8	22.5	79.4	22.1	77.1	16.6	88.2	17.8	89.5
CNN-GRU	11.7	94.8	19.3	<u>90.3</u>	<u>13.9</u>	89.9	13.7	91.4	<u>13.5</u>	<u>91.2</u>
Meta-Model (ours)	7.7	99.4	12.4	95.1	11.2	97.5	12.9	94.7	13.3	94.2

Best results appear in bold, while the second-best are marked with an underline

The results are obtained directly from the cited papers

seen from the results of Table 10 that the model continues to exhibit superior performance across the target domains. In particular, the model achieves the lowest Half Total Error Rate (HTER) in four protocols and the highest Area Under the Curve (AUC) score on all target domains. This outcome emphasizes that even with limited source data, the stacking ensemble approach maintains its robust generalization capability. Overall, this finding is significant as it demonstrates the effectiveness of the stacking ensemble approach in adapting to scenarios where a limited amount of source data is available.

#### 4.6 Ablation Studies

In order to assess the effectiveness of our proposed synthetic data generation and ensemble learning, we conducted separate experiments on four datasets in cross-testing scenarios. The results of these experiments are detailed in Table 11. The results without augmentation represent the performance of the deep DenseNet-201 model (Huang et al., 2017) on stabilized encoded training samples with alpha values set to 0.1. Subsequently, we gradually introduce synthetic training samples and assess the performance at different alpha values. Initially, with Synthetic set1 (alpha values = 0.5), the model demonstrates a slight improvement in performance, evident in a higher AUC and reduced HTER compared to training without augmentation. For instance, in the O&C&I to M scenario, the HTER decreased from 19.02 to 18.11, and the AUC increased from 86.12 to 87.63. Similarly, we assess the performance using Synthetic set2. It can be observed that the model continues to show enhanced performance.

Next, we assess the performance using Synthetic set2 (alpha values = 1.0). One can see that the model continues to show enhanced performance. The HTER improved by approximately 1.3% on the O&M&I to C scenario, reducing from 19.52 to 17.20. We then evaluate the third synthetic

subset, Synthetic set3 (alpha values = 1.5). This set shows consistent improvements across most scenarios. For example, in the O&C&I to M scenario, the HTER decreased to 17.21, and the AUC improved to 90.87. When combining all synthetic sets, the model achieves significant performance gains. The combined sets result in an overall reduction in HTER and an increase in AUC across all scenarios. For instance, the HTER in the O&M&I to C scenario drops to 15.20, and the AUC rises to 93.90. We also investigate the alpha values to 2.0, but this leads to a decline in the model's performance. Hence, we choose to report the performance for only three alpha values.

Finally, we assess the performance after introducing LSTM, BiLSTM, and GRU layers without using the metamodel. Each of these architectures further enhances the model's performance, for instance, CNN-LSTM achieves an HTER of 8.80 and AUC of 99.38 in the O&C&I to M scenario. CNN-BiLSTM represents an HTER of 2.33 and AUC of 99.98 in the O&M&I to C scenario. Similarly, CNN-GRU achieves an HTER of 8.04 and AUC of 94.02 in the O&M&I to I scenario. For the I&C&M to O scenario, CNN-BiLSTM shows an improved HTER of 12.28 and an AUC of 99.52, outperforming other RNNs.

#### 4.7 Discussion

The quantitative results in Table 11 highlight the critical role of synthetic images in training CNN models effectively. When the CNN is trained on all synthetic subsets, there is a notable improvement in performance across various datasets: up to 3% for M, 4% for C, 4% for I, and 6% for O. These gains are particularly significant given the challenging nature of cross-dataset scenarios. Motivated by these initial improvements, the study explores ensemble learning to fully leverage the variations present in synthetic data. Specifically, LSTM is integrated with CNN to predict temporal blur, which further

Table 11Ablation study usingcross-database evaluation

	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
Methods	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
CNN w/o augmentation	19.02	86.12	19.52	87.63	26.66	81.22	32.44	81.54
CNN + Synthetic set1	18.11	87.63	18.66	86.22	24.00	79.39	31.55	80.25
CNN + Synthetic set2	18.82	91.09	17.20	84.49	25.63	75.24	30.01	78.91
CNN + Synthetic set3	17.21	90.87	21.28	84.49	24.33	79.05	29.45	77.14
CNN + All sets	16.12	90.12	15.20	93.90	22.60	87.68	26.01	82.30
CNN-LSTM	8.80	99.38	3.26	99.95	15.52	97.62	14.31	98.11
CNN-BiLSTM	14.96	99.71	2.33	99.98	8.04	94.02	9.28	99.52
CNN-GRU	9.25	97.67	1.96	99.98	18.41	96.66	11.44	97.79
Meta model	1.20	99.98	1.37	99.99	5.40	99.97	2.08	99.77

enhances the model's performance across all datasets. This approach capitalizes on temporal information, allowing the model to better differentiate between genuine and spoofed images. Similarly, employing BiLSTM and GRU models results in substantial performance improvements, particularly notable on the C and O datasets.

However, the generalization ability of individual RNN models (LSTM, BiLSTM, and GRU) remains somewhat limited. To address this, the study introduces a meta-model designed to combine the strengths and mitigate the weaknesses of these base models. This meta-model is trained using the predictions of the base models, creating a new training set from these predictions. The meta-model, another RNN, learns to optimally weigh and combine the base model predictions, leading to enhanced overall performance. We argue that the ensemble learning guided by a video distillation scheme proves to be highly beneficial in improving crossdomain face PAD performance. In particular, the proposed data augmentation technique, which involves using synthetic images, facilitates the base models in learning from the diversity within the data. Temporal inconsistencies or blurriness, which might otherwise be considered noise, are leveraged as valuable features for distinguishing genuine images from spoofed ones. Thus, this approach ensures that the final test predictions made by the meta-model are robust and reliable.

#### 4.8 Comparisons of Execution Times

We analyze the execution times of the proposed video distillation technique with the previous global motion estimation methods (Muhammad et al., 2022a, b) and optical flow(Horn & Schunck, 1981). Table 12 reports the numerical results in the total number of seconds used to generate the training samples based on Eq. (4) using alpha values of 0.1 on two datasets. All these comparison results were reported by using a MATLAB environment based on a workstation with 3.5 GHz Intel Core i7-5930k and 64 GB RAM. One can see that the proposed global motion estimation technique is computationally less expensive than the previous motion estimated methods reported recently in the literature. This is due to the fact that the FAST (Rosten & Drummond, 2005) feature detector is designed with a focus on computational efficiency and speed, while FREAK (Alahi et al., 2012) is intended to work in combination with fast feature detectors like FAST, providing a matching mechanism that is both fast and robust.

#### 4.9 Interpretation of the Deep Neural Network

Interpretation is essential to observe the learning patterns in data that are important, but there is no clear consensus on how interpretability should be best defined in the context of machine learning (Molnar et al., 2020). Although explanation methods intend to make neural networks more trustworthy and interpretable, the "black-box" nature of deep neural networks can make it challenging to determine precisely why a particular decision was made. For instance, synthetic samples provide additional variations of the data and lead to better interpretability compared to the same model trained without synthetic samples. This improvement is attributed to the fact that the motion cues, which are naturally available in the frame sequences, are "easy to learn" for the model and play an important role in model optimization. Consequently, the importance of interpretation is becoming increasingly popular and has led to useful and promising findings.

In our work, we utilize Gradient-weighted Class Activation Mapping (denoted as Grad-CAM) (Selvaraju et al., 2017), Occlusion Sensitivity Maps (denoted as OCC-SEN) (Zeiler & Fergus, 2014), Gradient Attribution Map using Guided Backpropagation (denoted as Grad-ATT) (Springenberg et al., 2014), and Locally Interpretable Model-Agnostic Explanations (denoted as LIME) (Ribeiro et al., 2016) to understand which patterns in the data are deemed important and contribute to the final decision. These methods enable us to trust the behavior of the developed deep learning model and/or further tune the model by observing its interpretations. Specifically, we extract visualization maps from the

Table 12	Average	execution	time	in	seconds

Dataset	Optical flow (1	Horn & Schunck, 1981) ASGS method (Muhamma	d et al., 2022b) TSS method (Muhamm	ad et al., 2022a) Ours
CASIA-MFSD	1560	1487	1140	1023
REPLAY-ATTACK	1082	1003	780	641

Image Type	_	Grad-CAM	OCC-SEN	Grad-ATT	LIME	Masked
Encoded image			-		No.	
Synthetic sample 1					R	<b>\$</b>
Synthetic sample 2					<b>N</b>	<i>,</i> ≉_ <b>_</b> €
Synthetic sample 3						9 16
Encoded image				N.		# 10
Synthetic sample 1						
Synthetic sample 2						
Synthetic sample 3				<i>T</i>		

**Fig. 5** Displaying feature maps through visualization. The types of images are labelled in the first column. The second column shows the original encoded and synthetic images. The third column illustrates the feature maps from Grad-CAM (Selvaraju et al., 2017) while the fourth column shows the feature maps from occlusion sensitivity maps (Zeiler & Fergus, 2014). Similarly, the fifth and sixth column visu-

alize the features maps from Gradient Attribution map using Guided Backpropagation (Springenberg et al., 2014), and locally interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016), respectively. The last column shows the masked images obtained from LIME predictions pretrained DenseNet-201 (Huang et al., 2017) convolutional neural network for all of the above methods in our experiments. In Fig. 5, we visualize diverse synthetic images from the CASIA-FASD dataset. The first four rows show print attack images, while the next four rows show replay attack images. Each visualization method captures the class discriminative region, thanks to the proposed synthetic data generation scheme that allows the network to use more subtle cues for its correct classification. In particular, the first row shows that the neurons in the deep convolutional layers focus on the paper's texture and hand movement cues. However, Grad-ATT (Springenberg et al., 2014) interpretation shows that the model also takes the background as context to make the prediction. Surprisingly, this issue is eliminated by the proposed synthetic data generation scheme, where the second, third, and fourth rows show that the model only considers motion cues, surface edges, and barely observes the background context.

In the case of a replay attack, the remaining rows show that the tablet screen and hand movement provide discriminative information for the model's prediction. While we cannot present this for every image of the dataset, we observed that the mouth region, eye blinking, and head rotation contribute positively to distinguishing live and spoofed images. Thus, interpretations from the above methods demonstrate that the proposed learning model is focusing on the correct features of the input data, and the model's decision can be viewed in a human-understandable way. Moreover, the proposed synthetic data generation method provides informative RGB images and helps the model to make the features of spoofed faces more dispersed, allowing a better class discrimination to generalize well to the target domain.

#### 4.10 Visualization and Analysis

In order to visually illustrate the individual contributions of each model, we employ t-SNE (t-Distributed Stochastic Neighbor Embedding) to analyze the distribution of different features, as depicted in Fig. 6. Initially, the model is trained on the 0+C+I source domains without incorporating synthetic samples, resulting in a trivial distribution shown in Fig. 6a. In this representation, the boundary between live and spoofed samples is indistinct, and areas of overlap can lead to potential misclassifications, thereby degrading overall performance.

However, when synthetic samples are included in the model, as depicted in Fig. 6b, the feature distribution demonstrates improvement, offering a comparatively clearer separation compared to the model that does not include synthetic samples. This enhancement is attributed to the synthetic samples aiding the model in recognizing spatiotemporal artifacts.



Fig. 6 The t-SNE visualization of feature distributions on cross-testing scenarios. **a** shows the feature distribution of the original encoded video clips, **b** reflects the feature distribution of encoded video clips with a subset of synthetic samples, (c) shows the feature distribution of our meta-model



Fig. 7 The Receiver Operating Characteristics (ROC) curves. **a** O&C&I to M, **b** O&M&I to C, **c** O&C&M to I, and **d** I&C&M to O are developed by plotting the true positive rate (TPR) against the false positive rate (FPR)

Nonetheless, with the introduction of the meta-model, we observe a well-structured and compact distribution with a clearly defined boundary in Fig. 6c. As a result, our proposed ensemble learning approach exhibits strong generalizability when applied to unseen target data.

In Fig. 7, we employ ROC curves to visually represent the model's ability to differentiate between real and attack classes. As demonstrated in Fig. 7, the meta-model consistently achieves an AUC (Area Under the Curve) of over 90% across all datasets, showcasing an impressive level of performance on previously unseen testing sets. The ROC curve is constructed with True Positive Rate (TPR) plotted against False Positive Rate (FPR), where FPR is on the x-axis and TPR is on the y-axis. Specifically, when the meta-model (ensemble) shifts the curves closer to the top-left corner, it indicates superior performance in distinguishing between the classes.

### **5** Conclusions

In this paper, we addressed the domain generalization issue in face presentation attack detection (PAD) by proposing a novel approach that combines data augmentation and deep ensemble learning. By observing multiple blurred sequences, the base models were able to learn sequential patterns of motion blur and infer how objects move over time. This data augmentation technique was found to be helpful in improving the robustness of the models and their ability to handle variations in motion blur present in real-world scenarios. This technique not only enhanced the training data, but also addressed the issue of black framing that might arise during feature-based global motion estimation. Based on the experimental results, the performance of LSTM, BiL-STM, and GRU still face limitations in certain scenarios. To improve the overall generalization, we introduced a metamodel that leverages the strengths of different base models. This ensemble approach allowed the model to benefit from the diverse representations learned by individual base models. Based on the experimental results on seven benchmark datasets, the meta-model achieves competitive performance on all datasets.

Finally, the interpretation of the model shows that motion cues (e.g., temporal information or motion patterns) are helpful in improving the model's generalization ability. We conclude that the effectiveness of a meta-model depends on the diversity and quality of the base models used. Especially, if the base models suffer from similar limitations or biases, the meta-model may not provide significant improvements. Although the proposed method improves generalization, we observe two disadvantages. First, since the meta-model requires an additional training phase using the outputs from the base models, it introduces an extra layer of computational complexity and cost. Second, the proposed data augmentation may not encode all fine details, especially when the background is non-static. Therefore, future research should explore new approaches and enhancements to address these limitations and advance video summarization methods for non-static background videos.

Acknowledgements This work is financially supported by 'Understanding speech and scene with ears and eyes (USSEE)" (project number 345,791). The first author also acknowledges the support of the Ella and Georg Ehrnrooth foundation.

Funding Open Access funding provided by Aalto University.

**Data Availibility** The databases employed in our study are publicly available. The source code of the implementation performed in this paper is available on the project webpage: https://github.com/Usman1021/Ensemble.

#### Declarations

**Conflict of interest** The authors have no Conflict of interest that could have appeared to influence the work reported in this paper.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecomm ons.org/licenses/by/4.0/.

#### References

- Alahi, A., Ortiz, R., & Vandergheynst, P. (2012). Freak: Fast retina keypoint. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 510–517). IEEE.
- Antil, A., & Dhiman, C. (2024). Mf2shrt: Multi-modal feature fusion using shared layered transformer for face anti-spoofing. ACM Transactions on Multimedia Computing, Communications and Applications. https://doi.org/10.1145/3640817
- Arashloo, S. R. (2020). Unseen face presentation attack detection using sparse multiple kernel fisher null-space. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10), 4084–4095.
- Bharadwaj, S., Dhamecha, T.I., Vatsa, M., & Singh, R. (2013). Computationally efficient face spoofing detection with motion magnification. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 105–110).
- Boulkenafet, Z., Komulainen, J., Feng, X., & Hadid, A. (2016) Scale space texture analysis for face anti-spoofing. In 2016 international conference on biometrics (ICB) (pp. 1–6). IEEE.
- Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., & Hadid, A. (2017). Oulu-npu: A mobile face presentation attack database with realworld variations. In 2017 12th IEEE international conference on

*automatic face and gesture recognition (FG 2017)* (pp. 612–618). IEEE.

- Boulkenafet, Z., Komulainen, J., & Hadid, A. (2016). Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8), 1818–1830.
- Cai, R., Li, Z., Wan, R., Li, H., Hu, Y., & Kot, A. C. (2022). Learning meta pattern for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 17, 1201–1213.
- Chang, C.-J., Lee, Y.-C., Yao, S.-H., Chen, M.-H., Wang, C.-Y., Lai, S.-H., Chen, & T.P.-C. (2023). A closer look at geometric temporal dynamics for face anti-spoofing. In *Proceedings of the IEEE/CVF* conference on computer vision and pattern recognition (pp. 1081– 1091).
- Chingovska, I., Anjos, A., & Marcel, S. (2012) On the effectiveness of local binary patterns in face anti-spoofing. In 2012 BIOSIGproceedings of the international conference of biometrics special interest group (BIOSIG) (pp. 1–7). IEEE.
- Chuang, C.-C., Wang, C.-Y., & Lai, S.-H. (2023). Generalized face antispoofing via multi-task learning and one-side meta triplet loss. In 2023 IEEE 17th international conference on automatic face and gesture recognition (FG) (pp. 1–8). IEEE.
- Déniz, O., Bueno, G., Bermejo, E., & Sukthankar, R. (2011). Fast and accurate global motion compensation. *Pattern Recognition*, 44(12), 2887–2901.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv:2010.11929
- Fang, H., Liu, A., Jiang, N., Lu, Q., Zhao, G., & Wan, J. (2024). VI-fas: Domain generalization via vision-language model for face antispoofing. In *ICASSP 2024–2024 IEEE international conference* on acoustics, speech and signal processing (ICASSP) (pp. 4770– 4774). IEEE.
- Fatemifar, S., Awais, M., Akbari, A., & Kittler, J. (2020). A stacking ensemble for anomaly based client-specific face spoofing detection. In 2020 IEEE international conference on image processing (ICIP) (pp. 1371–1375). IEEE.
- Fatemifar, S., Awais, M., Arashloo, S.R., & Kittler, J. (2019). Combining multiple one-class classifiers for anomaly based face spoofing attack detection. In 2019 international conference on biometrics (ICB) (pp. 1–7). IEEE.
- Freitas Pereira, T.D., Anjos, A., Martino, J.M.D., & Marcel, S. (2012). Lbp-top based countermeasure against face spoofing attacks. In Asian conference on computer vision (pp. 121–132). Springer.
- Ganaie, M., Hu, M., et al. (2021). Ensemble deep learning: A review. arXiv:2104.02395
- George, A., & Marcel, S. (2019). Deep pixel-wise binary supervision for face presentation attack detection. In 2019 international conference on biometrics (ICB) (pp. 1–8). IEEE.
- George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., & Marcel, S. (2019). Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15, 42–55.
- Grother, P., Grother, P., Ngan, M., & Hanaoka, K. (2019). Face recognition vendor test (FRVT) part 2: Identification. US Department of Commerce, National Institute of Standards and Technology.
- Guo, X., Liu, Y., Jain, A., & Liu, X. (2022). Multi-domain learning for updating face anti-spoofing models. In *European conference on computer vision* (pp. 230–249). Springer.
- He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer* vision and pattern recognition (pp. 770–778).
- 🖄 Springer

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. Artificial Intelligence, 17(1–3), 185–203.
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K.Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).
- Huang, H.-P., Sun, D., Liu, Y., Chu, W.-S., Xiao, T., Yuan, J., Adam, H., & Yang, M.-H. (2022). Adaptive transformers for robust fewshot cross-domain face anti-spoofing. In *European conference on computer vision* (pp. 37–54). Springer.
- Huang, H., Xiang, Y., Yang, G., Lv, L., Li, X., Weng, Z., & Fu, Y. (2022). Generalized face anti-spoofing via cross-adversarial disentanglement with mixing augmentation. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal* processing (ICASSP) (pp. 2939–2943). IEEE.
- Jia, Y., Zhang, J., Shan, S., & Chen, X. (2020) Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF* conference on computer vision and pattern recognition (pp. 8484– 8493).
- Jia, S., Li, X., Hu, C., Guo, G., & Xu, Z. (2020). 3d face anti-spoofing with factorized bilinear coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10), 4031–4045.
- Jiang, J., & Sun, Y. (2022). Depth-based ensemble learning network for face anti-spoofing. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2954–2958). IEEE.
- Jia, Y., Zhang, J., Shan, S., & Chen, X. (2021). Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing. *Pattern Recognition*, 115, 107888.
- Kim, J. H., Jang, J., Kim, Y., & Nan, D. (2022). A structural topic model for exploring user satisfaction with mobile payments. *Computers, Materials and Continua*, 73(2), 3815–3826.
- Kwak, Y., Jung, M., Yoo, H., Shin, J., & Kim, C. (2023). Liveness score-based regression neural networks for face anti-spoofing. In ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1–5). IEEE.
- Li, L., & Feng, X. (2019). Face anti-spoofing via deep local binary pattern. In *Deep learning in object detection and recognition* (pp. 91–111).
- Liao, C.-H., Chen, W.-C., Liu, H.-T., Yeh, Y.-R., Hu, M.-C., & Chen, C.-S. (2023). Domain invariant vision transformer learning for face anti-spoofing. In *Proceedings of the IEEE/CVF winter conference* on applications of computer vision (pp. 6098–6107).
- Li, H., He, P., Wang, S., Rocha, A., Jiang, X., & Kot, A. C. (2018). Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10), 2639–2652.
- Liu, Y., Jourabloo, A., & Liu, X. (2018). Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of* the IEEE conference on computer vision and pattern recognition (pp. 389–398).
- Liu, Y., Stehouwer, J., & Liu, X. (2020) On disentangling spoof trace for generic face anti-spoofing. In *European conference on computer* vision (pp. 406–422). Springer.
- Liu, Y., Stehouwer, J., Jourabloo, A., & Liu, X. (2019). Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4680–4689).
- Liu, A., Tan, Z., Wan, J., Escalera, S., Guo, G., & Li, S.Z. (2021) 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 (pp. 1179–1187).

- Liu, A., Xue, S., Gan, J., Wan, J., Liang, Y., Deng, J., Escalera, S., & Lei, Z. (2024). Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. arXiv:2403.14333
- Liu, S.-Q., Lan, X., & Yuen, P. C. (2021). Multi-channel remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16, 2683–2696.
- Liu, Y., & Liu, X. (2022). Spoof trace disentanglement for generic face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), 3813–3830.
- Liu, S., Lu, S., Xu, H., Yang, J., Ding, S., & Ma, L. (2022). Feature generation and hypothesis verification for reliable face anti-spoofing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, 1782–1791.
- Liu, M., Mu, J., Yu, Z., Ruan, K., Shu, B., & Yang, J. (2022). Adversarial learning and decomposition-based domain generalization for face anti-spoofing. *Pattern Recognition Letters*, 155, 171–177.
- Liu, A., Tan, Z., Wan, J., Liang, Y., Lei, Z., Guo, G., & Li, S. Z. (2021). Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security*, 16, 2759–2772.
- Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning—A brief history, state-of-the-art and challenges. In *Joint European conference on machine learning and knowledge discov*ery in databases (pp. 417–431). Springer.
- Muhammad, U., & Hadid, A. (2019) Face anti-spoofing using hybrid residual learning framework. In 2019 international conference on biometrics (ICB) (pp. 1–7). IEEE.
- Muhammad, U., & Oussalah, M. (2023). Self-supervised face presentation attack detection with dynamic grayscale snippets. In 2023 IEEE 17th international conference on automatic face and gesture recognition (FG) (pp. 1–6). IEEE.
- Muhammad, U., Holmberg, T., Melo, W.C., & Hadid, A. (2019). Face anti-spoofing via sample learning based recurrent neural network (RNN). In *BMVC* (p. 113).
- Muhammad, U., & Oussalah, M. (2022). Face anti-spoofing from the perspective of data sampling. *Electronics Letters*, 59(1), e12692.
- Muhammad, U., & Oussalah, M. (2023). Face anti-spoofing from the perspective of data sampling. *Electronics Letters*, 59(1), 12692.
- Muhammad, U., Yu, Z., & Komulainen, J. (2022). Self-supervised 2d face presentation attack detection via temporal sequence sampling. *Pattern Recognition Letters*, 156, 15–22.
- Muhammad, U., Zhang, J., Liu, L., & Oussalah, M. (2022). An adaptive spatio-temporal global sampling for presentation attack detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*. https://doi.org/10.1109/TCSII.2022.3169435
- Parkin, A., & Grinchuk, O. (2019). Recognizing multi-modal face spoofing with face recognition networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 0–0).
- Patel, K., Han, H., & Jain, A. K. (2016). Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10), 2268–2283.
- Polikar, R. (2012). Ensemble learning. Ensemble machine learning: Methods and applications (pp. 1–34).
- Prechelt, L. (1998). Early stopping-but when? In *Neural networks: Tricks of the trade* (pp. 55–69). Springer.
- Qin, Y., Yu, Z., Yan, L., Wang, Z., Zhao, C., & Lei, Z. (2021). Meta-teacher for face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 6311–6326.
- Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). Why should i trust you? Explaining the predictions of any classifier. In *Proceedings* of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135–1144).

- Rosten, E., & Drummond, T. (2005). Fusing points and lines for high performance tracking. In *10th IEEE international conference on computer vision (ICCV'05)* (Vol.1, 2, pp. 1508–1515). IEEE.
- Saha, S., Xu, W., Kanakis, M., Georgoulis, S., Chen, Y., Paudel, D.P., & Van Gool, L. (2020). Domain agnostic feature learning for image and video based face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* workshops (pp. 802–803).
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673– 2681.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).
- Shahhosseini, M., Hu, G., & Pham, H. (2022). Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications*, 7, 100251.
- Shao, R., Lan, X., Li, J., & Yuen, P.C. (2019). Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10023–10031).
- Shao, R., Lan, X., & Yuen, P. C. (2020). Regularized fine-grained meta face anti-spoofing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, 11974–11981.
- Springenberg, J.T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. arXiv:1412.6806
- Srivatsan, K., Naseer, M., & Nandakumar, K. (2023). Flip: Crossdomain face anti-spoofing with language guidance. In *Proceedings* of the IEEE/CVF international conference on computer vision (pp. 19685–19696).
- Sun, Y., Liu, Y., Liu, X., Li, Y., & Chu, W.-S. (2023). Rethinking domain generalization for face anti-spoofing: Separability and alignment. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 24563–24574).
- Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Advances in neural information processing systems, Vol. 30.
- Tirunagari, S., Poh, N., Windridge, D., Iorliam, A., Suki, N., & Ho, A. T. (2015). Detection of face spoofing using visual dynamics. *IEEE Transactions on Information Forensics and Security*, 10(4), 762–777.
- Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., & Uszkoreit, J. (2021). MLP-mixer: An all-MLP architecture for vision. Advances in Neural Information Processing Systems, 34, 24261–24272.
- Torr, P. H., & Zisserman, A. (1997). Robust parameterization and computation of the trifocal tensor. *Image and Vision Computing*, 15(8), 591–605.
- Vareto, R. H., & Schwartz, W. R. (2021). Face spoofing detection via ensemble of classifiers toward low-power devices. *Pattern Analysis* and Applications, 24(2), 511–521.
- Wang, G., Han, H., Shan, S., Chen, X. (2020) Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF conference* on computer vision and pattern recognition (pp. 6678–6687).
- Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., & Wang, Z. (2022). Domain generalization via shuffled style assembly for face antispoofing. In *Proceedings of the IEEE/CVF conference on computer* vision and pattern recognition (pp. 4123–4133).
- Wang, Z., Yu, Z., Wang, X., Qin, Y., Li, J., Zhao, C., Lei, Z., Liu, X., Li, S., & Wang, Z. (2021). Consistency regularization for deep face anti-spoofing. arXiv:2111.12320

- Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., Zhou, F., & Lei, Z. (2020). Deep spatial gradient and temporal depth learning for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5042–5051).
- Wang, W., Wen, F., Zheng, H., Ying, R., & Liu, P. (2022). Conv-MLP: A convolution and MLP mixed model for multimodal face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 17, 2284–2297.
- Wen, D., Han, H., & Jain, A. K. (2015). Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics* and Security, 10(4), 746–761.
- Wu, H., Zeng, D., Hu, Y., Shi, H., & Mei, T. (2021). Dual spoof disentanglement generation for face anti-spoofing with depth uncertainty learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7), 4626–4638.
- Yang, J., Lei, Z., Liao, S., & Li, S.Z. (2013). Face liveness detection with component dependent descriptor. In 2013 international conference on biometrics (ICB) (pp. 1–6). IEEE.
- Yin, W., Ming, Y., & Tian, L. (2016). A face anti-spoofing method based on optical flow field. In 2016 IEEE 13th international conference on signal processing (ICSP) (pp. 1333–1337). IEEE.
- Yu, Z., Cai, R., Cui, Y., Liu, X., Hu, Y., Kot, A. (2023). Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. arXiv:2302.05744
- Yu, Z., Qin, Y., Li, X., Wang, Z., Zhao, C., Lei, Z., & Zhao, G. (2020). Multi-modal face anti-spoofing based on central difference networks. In *Proceedings of the IEEE/CVF conference on computer* vision and pattern recognition workshops (pp. 650–651).
- Yue, H., Wang, K., Zhang, G., Feng, H., Han, J., Ding, E., & Wang, J. (2023). Cyclically disentangled feature translation for face anti-spoofing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37, 3358–3366.
- Zeiler, M.D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.
- Zhang, S., Wang, X., Liu, A., Zhao, C., Wan, J., Escalera, S., Shi, H., Wang, Z., & Li, S.Z. (2019) A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 919–928).

- Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., & Li, S.Z. (2012). A face antispoofing database with diverse attacks. In 2012 5th IAPR international conference on biometrics (ICB) (pp. 26–31). IEEE.
- Zhang, S., Liu, A., Wan, J., Liang, Y., Guo, G., Escalera, S., Escalante, H. J., & Li, S. Z. (2020). Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science, 2*(2), 182–193.
- Zhang, N., Liu, W., & Xia, X. (2023). Video global motion compensation based on affine inverse transform model. *Sensors*, 23(18), 7750.
- Zheng, G., Liu, Y., Dai, W., Li, C., Zou, J., & Xiong, H. (2023) Learning causal representations for generalizable face anti spoofing. In ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1–5). IEEE.
- Zhou, Q., Zhang, K.-Y., Yao, T., Lu, X., Yi, R., Ding, S., & Ma, L. (2023). Instance-aware domain generalization for face antispoofing. In *Proceedings of the IEEE/CVF conference on computer* vision and pattern recognition (pp. 20453–20463).
- Zhou, Q., Zhang, K.-Y., Yao, T., Yi, R., Sheng, K., Ding, S., & Ma, L. (2022). Generative domain adaptation for face anti-spoofing. In *European conference on computer vision* (pp. 335–356). Springer.
- Zhou, K., Yang, Y., Qiao, Y., & Xiang, T. (2021). Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30, 8008–8018.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.