2022 26th International Conference on Pattern Recognition (ICPR)
August 21-25, 2022, Montréal, Québec, Canada

Disentangling Facial Pose and Appearance Information for Face Anti-spoofing

Ajian Liu
Macau University of Science and Technology
Institute of Automation Chinese Academy of Sciences
Beijing, China
Email: ajianliu92@gmail.com

Jun Wan
Macau University of Science and Technology
Institute of Automation Chinese Academy of Sciences
Beijing, China
Email: jun.wan@nlpr.ia.ac.cn

Ning Jiang
Mashang Consumer Finance Co., Ltd.
Beijing, China
Email: ning.jiang02@msxf.com

Hongbin wang
Mashang Consumer Finance Co., Ltd.
Beijing, China
Email: hongbin.wang02@msxf.com

Yanyan Liang
Macau University of Science and Technology
Macau, China
Email: yyliang@must.edu.mo

2022 26th International Conference on Pattern Recognition (ICPR) | 978-1-6654-9062-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPR56361.2022.9956674

***Abstract*—Face Anti-spoofing aims to determine whether the captured face from a face recognition system is real or fake. However, the facial pose and local significant spoofing traces (*i.e.*, the boundary and reflection spot in presentation attack instruments) seriously affects the performance and stability of the current algorithms. Due to they regard the face image as an indivisible unit, and process it holistically, rarely consider excluding these liveness-irrelated factors. Unlike it, we design a Pose-Independent Face Anti-Spoofing (PIFAS) framework to disentangle face into an appearance information and a pose code to capture liveness and liveness-irrelated features, respectively. Specifically, the PIFAS consists of an Unsupervised Pose Switching (UPS) module and a Mutual Information Averaged Defense (MIAD) module, which are used to control the facial pose and suppress the local significant attack traces by averaging the local and global knowledge. Extensive experimental evaluations on multiple face anti-spoofing datasets verify that the proposed method can improve the generalization and stabilize the performance of each testing video through alleviating the interference from liveness-irrelated factors.**

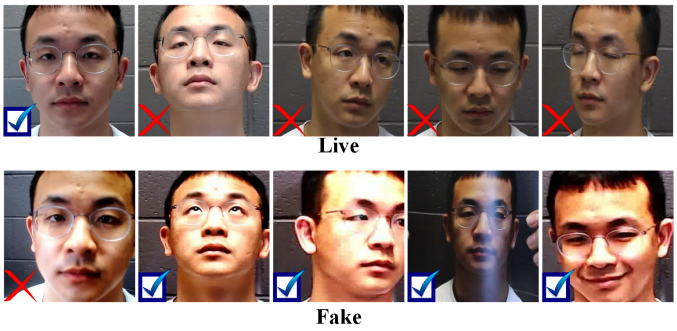

Fig. 1. Same detection model [11] makes different judgments for different frames in one testing video. These samples are drawn from SiW dataset. The 'Red Fork' indicates that the model misclassifies the samples. For example, the 'Red Fork' in the first line means that the live face is misclassified into the attack face, and the attack face in the second line is misclassified into the live face. The 'Blue checkbox' indicates that the model classifies the samples correctly.

I. Introduction

Face anti-spoofing (FAS) is critical to prevent the face recognition system from malicious attacks, such as print attack [1], replay attack [2], or 3D attacks [3]. It has become an increasingly concerns [4], [5], [6], [7], [8], [9], [10] recently due to the widespread application of face recognition in financial payment, access control, and phone unlocking.

Some early CNN based Presentation Attack Detection (PAD) methods [12], [13], [14] regard the FAS as a binary classification task. It might discover arbitrary clues that can separate the two classes (live or fake), such as facial pose, local bright spot, and screen bezel, but not the faithful spoofing patterns [11]. Inspired by this, some recent works [11], [4], [15] leverage the physical-based depth information instead of binary softmax loss as supervision, and aim to predict the true depth of the faces with the supervision of the depth maps and flat masks. Although these methods achieve good performances in many benchmarks [16], [11], [17], [18], [19], [20], [21], there are still some shortcomings. For example, one model is extremely sensitive to the facial poses and significant local spoofing traces. It is difficult to meet the actual deployment of a FAS system, especially when the model judges different frames in the same video as different categories. The predicted stability among these methods are rarely considered. As shown in Fig. 1, for live sequences, the model [11] may mistakenly live frame as fake ones based on local reflected light spots, shadows or pose, while for fake sequences, the model misclassifies the fake frame as live ones due to these clues were not detected. Therefore, how to effectively eliminate the interference of the facial poses and significant local spoofing traces is an effective strategy to improve these algorithms.

There are two limitations that lead to the defects of the above model. (1) From the perspective of face composition, almost all the prior works process the face image as an independent unit, and ignore subtle spoofing clues that are highly susceptible to the facial poses. (2) From the perspective

of face structure, almost all the prior works process the face image as a complete unit, and lack the consideration of the relationship between local patches and global face. It induces the model to pay much attention to some significant local spoofing traces, and ignore the other paths or global characteristics of fake faces during the training time.

For the first limitation, one possible solution is to separate the face into different components by the disentangled representation learning [22], [23]. Zhang *et al.* [6] disentangle latent space of face into two sub-spaces: liveness space and content space that integrates remaining liveness-irrelated information. Other works disentangle the spoofing clues from the input faces, such as the noise patterns [24], [25] or spoofing traces [26], and these clues is further used for classification. However, on the one hand, these liveness features or spoofing clues are so subtle, and difficult to peel off from the face image. On the other hand, they are sensitive to facial poses, and leading to unstable performance even for the same testing video. Motivated by [27], [28], we disentangle the face into an appearance information and a pose code, and randomly replace the facial pose in an unsupervised manner to alleviate the model's bias. For the second limitation, one reasonable solution is to split the facial images into some local subsets, and to learn the liveness features by exploring relationships between the global and local subsets. Thus the model's bias to some local spoofing clues is alleviated under the guidance of category consistency. Prior work [29] has explored the fusing patch-based and holistic depth-based clues for extracting the local features and global depth maps. However, it neither studies the interaction between local and global representations, or does it make good use of the auxiliary supervision that the category consistency of each locality and complete face. In fact, global representations plays a stronger guiding role for the model's attention on effective liveness features instead of local arbitrary clues. Inspired by [30], we split the global representations of facial images into some local subsets, and learn the liveness features by averaging mutual information between local and global representations.

To sum up, the contributions are summarized as follows:

- We propose a simple yet effective framework, namely Pose-independent Face Anti-spoofing (PIFAS), against the inferences of liveness-irrelated factors by disentangling face into appearance information and pose codes.
- An Unsupervised Pose Switching (UPS) module is introduced as the first stage in the PIFAS. It completes the facial pose replacement through a generative way under the guidance of facial geometric maps.
- A Mutual Information Averaged Defense (MIAD) module is adopted as the second stage in the PIFAS. It incorporates knowledge about locality in the face into a score map under the category consistency with the global input.
- Extensive experiments demonstrate that the proposed method achieves competitive results on several benchmarks, especially for stabilizing the performance without being affected by liveness-irrelated factors.

II. Proposed Method

A. Pose-independent Face Anti-spoofing

As shown in Fig. 1, we analyze that the bias to liveness-irrelated factors (such as facial pose) and the overfitting to some local spoofing clues (such as local reflected light spots and shadows) are two main reasons for the unstable performance of the current algorithms. In order to tackle the above problems, as shown in Fig. 2, we propose a **PIFAS** framework with two stages: an Unsupervised Pose Switching (**UPS**) stage (light green) and a Mutual Information Averaged Defense (**MIAD**) stage (light yellow).

Specially, in the first stage for a mini-batch, any input sample (whether live or fake face, denoted as $\mathbf{S} \in \{\mathbf{S^l} \cup \mathbf{S^f}\}$) is arrived at where it started with an unsupervised manner by being disentangled through **Enc-A** module (appearance encoder E^a) and reconstructed through **Dec** module (face decoder De). Simultaneously, the according geometric map (denoted as $\mathbf{D}$) of each sample is utilized as the pose guidance in **Enc-P** module (pose encoder E^p) to tell Enc-A what liveness-irrelated factor is and to disentangle it from liveness features. After the first stage, any input face $\mathbf{S}$ will generate a counterpart that is consistent with its category but has a different facial pose, denoted as $\hat{\mathbf{S}}$.

While in the second stage, inspired by [30], we first encode the image $\mathbf{S}$ (or $\hat{\mathbf{S}}$) to a representation map of $M \times M$ feature vectors corresponding to $M \times M$ local patches through **Enc-R** module (representation extractor E^r). We further summarize this representation map into a global feature vector $\mathbf{G}$, and add $\mathbf{G}$ with the lower-level feature $\mathbf{L}(\mathbf{i}, \mathbf{j})$ at every location (represents position coordinate) to form a local-global feature pair. Subsequently, we calculate a score for each local-global pair through a **Depth Estimator** (abbreviated as Dep). After the second stage, any input face will generate a score map with size of $M \times M$ that used to decide its final category under the supervision of a pre-defined feature-level label.

B. Unsupervised Facial Pose Switching

Directly disentangling facial pose from a given face is a sub-optimal choice, since it may be still entangled with liveness features [6]. Therefore, how to accurately describe the facial pose and completely disentangle it are the two main tasks of UPS model. Inspired by recent 3D reconstruction methods [31] which are widely used to estimate the pseudo-depth information in face anti-spoofing, we find that they are also very accurate in portraying facial poses [28].

In this work, we disentangle face into an appearance information and a pose code to capture the liveness features and reflect the liveness-irrelated factor respectively, where the 3D geometric maps that are utilized as the pose guidance for accurately describing the facial pose. As shown in Fig. 2, our UPS module consists of an appearance encoder E^a, a pose encoder E^p, which extract the appearance information $E^a(\mathbf{S})$ and pose code $E^p(\mathbf{D})$, respectively. As for the decoder De, inspired by recent works [32], [27] that use affine transformation parameters in normalization layers to represent

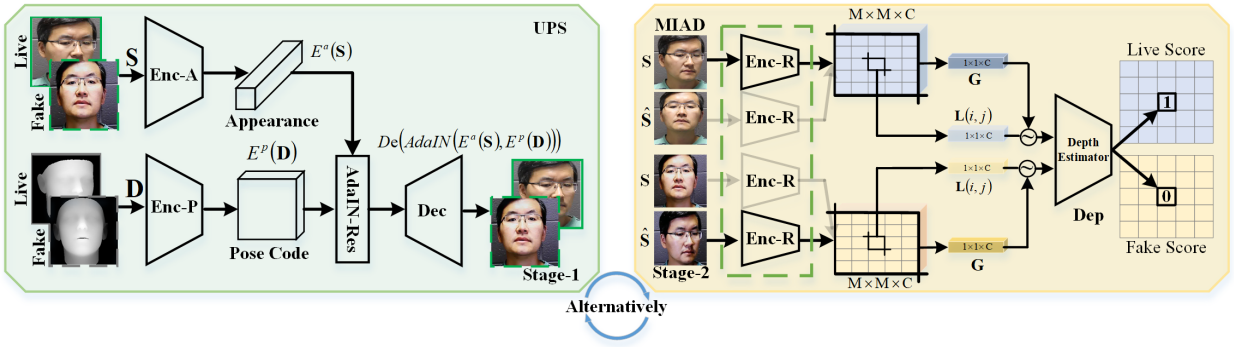

Fig. 2. The overall architecture of the Pose-independent Face Anti-spoofing (PIFAS). It is completed in two stages: an Unsupervised Pose Switching (UPS) stage (light green) and a Mutual Information Averaged Defense (MIAD) stage (light yellow). The former aims to disentangle the facial pose from the whole face, and replace it with agnostic poses randomly to alleviate the bias of the defense model on specific facial poses. And the latter aims to alleviate the bias of defense model on local spoofing traces by suppressing the significant effects of its representations. Note that all Enc-R modules that in the light green box share weights.

styles, we equip the residual blocks with Adaptive Instance Normalization (AdaINRes) [32] layers whose parameters are dynamically generated by a multilayer perceptron (MLP) from the appearance information. Therefore, our decoder reconstructs the original input $\mathbf{S}$ from its pose code $E^p(\mathbf{D})$ and appearance information $E^a(\mathbf{S})$. Since no supervision is used in the process, our UPS module is trained in an unsupervised way.

Given an image $\mathbf{S}$ with its 3D geometric map $\mathbf{D}$ form a mini-batch, we should be able to reconstruct it after encoding by E^a, E^p and decoding by De, sequentially. Such as: $\mathbf{S}/\mathbf{D} \rightarrow E^a(\mathbf{S})/E^p(\mathbf{D}) \rightarrow De(AdaINRes(E^a(\mathbf{S}), E^p(\mathbf{D}))) = \hat{\mathbf{S}} \approx \mathbf{S}$.

$$\mathcal{L}_{Recon}^{image} = \mathbb{E}_{\mathbf{S},\hat{\mathbf{S}}}\left[\left\|\hat{\mathbf{S}} - \mathbf{S}\right\|_1\right] \tag{1}$$

where $\mathbf{S}$ and $\hat{\mathbf{S}}$ have the same pose at this time. While given latent codes, such as appearance information and pose code which are encoded from E^a and E^p at translation time, we should be able to reconstruct them after decoding by De and encoding by E^a and E^p again.

$$\mathcal{L}_{Recon}^{appearance} = \mathbb{E}_{\mathbf{S},\hat{\mathbf{S}}}\left[\left\|E^a(\hat{\mathbf{S}}) - E^a(\mathbf{S})\right\|_1\right] \tag{2}$$

$$\mathcal{L}_{Recon}^{pose} = \mathbb{E}_{\hat{\mathbf{S}},\mathbf{D}}\left[\left\|E^p(\hat{\mathbf{S}}) - E^p(\mathbf{D})\right\|_1\right] \tag{3}$$

In order to make the reconstructed sample $\hat{\mathbf{S}}$ by our UPS module that is indistinguishable from input sample $\mathbf{S}$, we employ GANs to align their distribution at the image level.

$$\mathcal{L}_{GAN}^{Image} = \mathbb{E}_{\mathbf{S}}\left[logDis^m(\mathbf{S})\right] + \mathbb{E}_{\hat{\mathbf{S}}}\left[log(1 - Dis^m(\hat{\mathbf{S}}))\right] \tag{4}$$

where Dis^m is multi-scale discriminator that tries to distinguish between generated image $\hat{\mathbf{S}}$ and original image in $\mathbf{S}$. Similar to [27], we train the encoders (E^a and E^p), decoder (De), and discriminator (Dis^m) to optimize the final objective for UPS module, which is a weighted sum of the reconstruction loss terms and adversarial loss.

$$\mathcal{L}_{UPS} = \lambda_1\left(\mathcal{L}_{Recon}^{image} + \mathcal{L}_{Recon}^{appearance} + \mathcal{L}_{Recon}^{pose}\right) + \mathcal{L}_{GAN}^{Image} \tag{5}$$

C. Mutual Information Averaged Defense

How to incorporate knowledge about locality in the input into the global feature and determine the classification score are the two main tasks of MIAD module.

In this work, instead of maximizing mutual information [30] between a local input and the output from a deep neural network encoder, we average the mutual information for monitoring the unexpected local spoofing clues under the reference of global representation. Then, we further control score distributions of the averaged local-global representation by matching to a pre-defined feature-level map. Specifically, we use the feature-level supervision by computing the probability distribution for each local-global pair through a depth estimator.

Given an image $\mathbf{S}$ with its one counterpart $\hat{\mathbf{S}}$ that is consistent with its pose but has a different category, we will get their score map by passing both the global-level feature vector ($\mathbf{G}$), and the local-level $M \times M$ feature map ($\mathbf{L}$), through a depth estimator Dep, *i.e.*, $\mathbf{S} \rightarrow E^r(\mathbf{S}) \rightarrow G \sim L(i,j) \rightarrow Dep(.) \rightarrow map_{(i,j)}$, '$\sim$' means averaging operation. The score map for $\hat{\mathbf{S}}$ is obtained similarly. The training of MIAD is completed under the supervision of pre-provided feature-level label. For live faces, map should be 1, and for fake faces as well as synthesized fake faces, map should be 0. We apply the $\mathcal{L}_2$ norm on this loss as:

$$\mathcal{L}_{MIAD} = \mathbb{E}_{\mathbf{S}\in\mathbf{S^l}}\left\|map_{\mathbf{S}} - 1\right\|_2^2 + \mathbb{E}_{\hat{\mathbf{S}}\in\mathbf{S^f}}\left\|map_{\hat{\mathbf{S}}}\right\|_2^2 \tag{6}$$

where $\mathbf{S^l}$, $\mathbf{S^f}$ are sample sets of live and fake faces, respectively. In the testing phase, we use the average of the output from map for classification:

$$score = 1/M^2\left\|map\right\|_1 \tag{7}$$

where M is the size of map which ideally is a scale within $(0, 1]$ for live faces and 0 for fake faces. we calculate the score by normalizing the norm ($\mathcal{L}_1$ norm) of fitted score map, and then judge the class of the testing face based on a given threshold.

III. EXPERIMENTS

In this section, we conduct a series of experiments on three widely used face anti-spoofing datasets, including OULU-NPU [16], SiW [11], and CASIA-SURF CeFA (briefly named CeFA) [33], to visually and quantitatively demonstrate the effectiveness of the proposed approach.

A. Experimental Setup

Datasets & Protocols. OULU-NPU [16] is a high-resolution dataset, consisting of $4,950$ real access and spoofing videos with many real-world variations. It contains 4 protocols to assess the effect of methods in one previously unseen condition. Similar, **SiW** [11] defines 3 protocols by introducing another three unknown testing conditions. **CeFA** [33] is a cross-ethnicity face anti-spoofing dataset, covering 3 ethnicities, 3 modalities, $1,607$ subjects. It consists of print, video-replay, and 3D mask attacks. Three protocols are reported in our experiments according to the official definition. It is worth mentioning that SiW and CeFA have subjects with much variations in poses, illuminations, expressions (PIE), which are more suitable for us to study how to improve the performance instability caused by changes in facial pose and lighting environment.

Evaluation metrics. For a fair comparison with prior methods, the following metrics are used in experimental results. Especially, Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), Average Classification Error Rate (ACER) describe the performance given a predetermined threshold, which is used for the metrics of OULU-NPU, SiW, and CeFA. In which the decision threshold is found from the development sets by minimizing the Equal Error Rate (EER).

B. Implementation Details

Training details. The proposed framework is implemented on a single NVIDIA TITAN X GPU. We resize the cropped face region to 256×256. In the training stage, all models are trained with a batch seize of 2 (one live and one fake face) and an initial learning rate of 0.0001. We train models with 40 epochs from scratch via Adam solver, and keep the same learning rate for the first 20 epochs and linearly decay it to 0 over the next 20 epochs. The λ_1 in Eq.5 is set to 10 according to the suggestion in [27]. The size M of the representation map in UPS module is set to 32, and the number of channel C for global and local features is 384. In the testing stage, only the feature extractor E^r and depth estimator Dep are used to inference and final score calculate in Eq.7.

Network Architecture. The UPS module consists of a generator and a multi-scale discriminator Dis^m with the same backbone with MUNIT [27]. In which the generator consists of two encoders E^a and E^p, a decoder De. For the feature extractor E^r in MIAD module, we employ the same architecture with Aux.(Depth) [11]. Our depth estimator Dep contains 3 convolutions to estimate the depth map by reducing the channel from 384 to 128, 64, and finally to 1.

TABLE I
EVALUATION RESULTS ON FOUR PROTOCOLS OF OULU-NPU.

P.	Method	APCER(%)	BPCER(%)	ACER(%)
1	STASN [34]	1.2	2.5	1.9
	Auxiliary [11]	1.6	1.6	1.6
	STDN [26]	0.8	**1.3**	1.1
	CDCN [5]	**0.4**	1.7	**1.0**
	PIFAS	1.0	1.7	1.4
2	Auxiliary [11]	2.7	2.7	2.7
	STASN [34]	4.2	**0.3**	2.2
	STDN [34]	2.3	1.6	1.9
	CDCN [5]	1.5	1.4	1.5
	PIFAS	**1.1**	1.1	**1.1**
3	Auxiliary [11]	2.7±1.3	3.1±1.7	2.9±1.5
	STASN [34]	4.7±3.9	**0.9±1.2**	2.8±1.6
	STDN [34]	1.6±1.6	4.0±5.4	2.8±3.3
	CDCN [5]	2.4±1.3	2.2±2.0	2.3±1.4
	PIFAS	**1.5±1.7**	2.2±1.2	**2.0±2.2**
4	Auxiliary [11]	9.3±5.6	10.4±6.0	9.5±6.0
	STASN [34]	6.7±10.6	8.3±8.4	7.5±4.7
	STDN [34]	**2.3±3.6**	5.2±5.4	3.8±4.2
	CDCN [5]	4.6±4.6	9.2±8.0	6.9±2.9
	PIFAS	2.6±1.4	**4.4±2.4**	**3.5±2.7**

C. Experimental Comparison

Results on OULU-NPU. We compare the results of four protocols provided by OULU-NPU with four competitive methods, *i.e.*, Auxiliary [11], STASN [34], STDN [26] and CDCN [5].

As shown in Tab. I, our PIFAS achieves the best performance on Protocol 2, 3 and 4, respectively. Such as ACER values are 1.1%, 2.0%, and 3.5%. Since the collection environment of the OULU-NPU is relatively harmonious, compared with the baseline method Auxiliary [11], the performance improvement of our approach is not very obvious, such as the ACER is reduced by 0.2%, and 0.9 in Protocol 1, and 3 respectively. However, our method achieves significant improvements compared to Auxiliary in protocols 2 and 4, *i.e.*, the ACER is reduced by 1.6%, and 6.0%, respectively.

In fact, the Protocol 1 and Protocol 3 [16] introduce various image domains by setting up multiple acquisition sessions (illumination variation) and devices (camera variation). While the Protocol 2 and Protocol 4 mainly explore the generalization of the algorithm against unknown attacks. Compared with the baseline method, we attribute performance improvement mainly benefits from the UPS module that introduces two high-quality counterparts for each input during the training process: one is with different pose but same category, and the other is with different category but same pose, which alleviate the model's overfitting by enriching the diversity of face poses and samples.

Results on SiW. SiW brings great challenges to the face anti-spoofing task due to it covers much larger variations in facial poses, illuminations, expressions, and other practical factors. Tab. II lists the comparison results for the defined three protocols with five competitive methods, including Auxiliary [11], STASN [34], MetaFAS-DR [35], STDN [26], and CDCN [5].

TABLE II
EVALUATION RESULTS ON THREE PROTOCOLS OF SIW DATASET.

P.	Method	APCER(%)	BPCER(%)	ACER(%)
1	Auxiliary [11]	3.58	3.58	3.58
	STASN [34]	-	-	1.00
	MetaFAS-DR	0.52	0.50	0.51
	STDN	**0.00**	**0.00**	**0.00**
	CDCN [5]	0.07	0.17	0.12
	PIFAS	**0.00**	**0.00**	**0.00**
2	Auxiliary [11]	0.57±0.69	0.57±0.69	0.57±0.69
	MetaFAS-DR	0.25±0.32	0.33±0.27	0.29±0.28
	STASN [34]	-	-	0.28±0.05
	STDN	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
	CDCN [5]	0.00±0.00	0.13±0.09	0.06±0.04
	PIFAS	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
3	STASN [34]	-	-	12.10±1.50
	Auxiliary [11]	8.31±3.81	8.31±3.81	8.31±3.81
	STDN	8.30±3.30	7.50±3.30	7.90±3.30
	CDCN [5]	**1.67±0.11**	**1.76±0.12**	**1.71±0.11**
	PIFAS	5.18±3.62	7.68±2.31	6.43±3.25

Overall, our approach achieves the lowest ACER on Protocol 1, 2, respectively. For example, the proposed method outperforms the baseline method Auxiliary [11] with a significantly margin on Protocol 1 that deals with variations in face pose and expression, *i.e.*, APCER, BPCER, and ACER are all reduced by 3.58%.

In Protocol 3, although our method failed to achieve optimal performance, it still achieved a considerable improvement compared to the baseline method Auxiliary [11]. Such as the ACER is reduced from 8.31% to 6.43%. Those improvements mainly benefit from two aspects: (1) At the sample input level, the UPS module that can generate a lot of counterparts of being aligned with input faces, which prompts the model to focus on spoofing clues in the face region rather than changes in facial posture and expression. (2) At the feature learning level, the MIAD module incorporates knowledge about locality in the face with the global features to alleviate the bias of defense model on local spoofing clues by suppressing its significant effects. Finally, we analyze that the our approach is worse than CDCN [5] on Protocol 3 for it uses central difference convolution to extract features, which is more suitable for the environment with unknown attack type.

Results on CeFA. CeFA is another large-scale dataset with much larger variations in facial poses. We only conduct experiments on Protocol 1, 2, 4 with RGB modality, which are related to this work for the research points. The benchmark method SD-Net [33] and two competitive methods are compared with our approach in Tab. III, *i.e.*, BOBO [36] and VisionLabs [37]. Note that these two methods only report the experimental results on Protocol 4 used in the Chalearn CeFA Face Anti-Spoofing challenge [38].

It can be seen that our approach achieves the lowest ACER in Protocol 1, 2, and second lowest ACER on Protocol 4, which are 4.6%, 3.1%, and 5.9%, respectively. Concretely, our approach outperforms the benchmark results by a great advantage in the three protocols, *i.e.*, ACER reduces from 14.1%, 23.3% and 35.2% to 4.6%, 3.1% and 5.9%, respectively. We analyze that the our approach is worse than VisionLabs [37] on Protocol 4 for two reasons: (1) VisionLabs is based on the embedding integration of 4 parallel networks and extracts the 4 kinds of optical flow information of the sample respectively. (2) The result of VisionLabs is the best one obtained by submitting testing results to the competition website multiple times.

TABLE III
EVALUATION RESULTS ON THREE PROTOCOLS OF CEFA DATASET.

P.	Method	APCER(%)	BPCER(%)	ACER(%)
1	SD-Net [33]	15.7±5.3	12.4±2.2	14.1±3.8
	PIFAS	**3.7±1.7**	**5.5±2.3**	**4.6±2.2**
2	SD-Net [33]	45.0±39.1	**1.6±1.9**	23.3±18.6
	PIFAS	**2.6±1.8**	3.6±1.2	**3.1±2.5**
4	SD-Net [33]	65.8±16.4	8.3±6.5	35.2±5.8
	BOBO [36]	7.2±3.7	2.5±0.5	4.8±1.8
	VisionLabs [37]	**0.1±0.1**	5.3±2.4	**2.7±1.2**
	PIFAS	9.7±2.3	**2.1±1.2**	5.9±1.5

TABLE IV
CROSS TESTING ON CASIA-FASD VS. REPLAY-ATTACK.

Method	Train	Test	Train	Test
	CASIA-FASD	Replay-Attack	Replay-Attack	CASIA-FASD
Auxiliary [11]	27.6		28.4	
CDCN [5]	15.5		32.6	
PIFAS	**14.3**		**26.2**	

By summarizing the above experimental results, they are consistent with the conclusion on OULU-NPU and SiW datasets that our method relieves the bias of the model on facial pose and local spoofing cues from two levels.

D. Cross-dataset Testing

In this experiment, there are two cross-dataset testing protocols, one is that training on the CASIA-FASD [1] and testing on Replay-Attack [2], the second one is exchanging the training dataset and the testing dataset.

See from the Tab. IV, our proposed method reduces the cross-testing errors on the Replay-Attack and CASIA-FASD by 13.3% and 2.2% relative to Auxiliary [11], and by 1.2% and 6.4% relative to CDCN [5], respectively. It shows that our approach has superior generalization to unknown environments (domain). We find that although CDCN is superior for the unknown attack type in Tab. II, its performance in the unknown data domain is worse than that of the benchmark method Auxiliary. While, our approach outperforms the Auxiliary in both cases. We analyze that the proposed method generates face samples with diverse appearance, which improves the generalization performance of the model.

E. Ablation Study

In order to evaluate the contribution of each component in our framework, we introduce two variations according to the improvements, *i.e.*, the backbone Aux.(Depth) (briefly named

TABLE V
QUANTITATIVE ABLATION STUDY OF THE EACH COMPONENT.

Method	SiW(Protocol 1)			CeFA(Protocol 4)		
	APCER	BPCER	ACER	APCER	BPCER	ACER
Aux.D [11]	3.58	3.58	3.58	**5.9±4.3**	12.3±5.2	9.1±3.4
Aux.D w/ MIAD	2.34	2.18	2.26	7.4±3.4	5.6±4.3	6.5±2.7
Aux.D w/ UPS	1.10	1.01	1.05	9.5±4.2	5.3±3.1	7.4±3.6
PIFAS	**0.00**	**0.00**	**0.00**	9.7±2.3	**2.1±1.2**	**5.9±1.5**

Aux.D) [11] with MIAD (denoted as Aux.D $w/$ MIDAD), and Aux.D with UPS (denoted as Aux.D $w/$ UPS), and perform ablation study on the Protocol 1 of the SiW and Protocol 4 of the CeFA dataset. In which the "w" is the abbreviation of "with".

Effect of the UPS Module. From the results of SiW dataset in Tab. V, we can obverse that the most significant contribution comes from the term of UPS since the performance of Aux.D drops sharply. Such as, the results of ACER is reduced by 2.53%, and reduced by 1.32% when equipped with UPS and MIAD, respectively. It indicates that the UPS has the ability to cope with changes in facial poses and expressions. Finally, if UPS and MIAD modules are equipped on the backbone at the same time, the performance will be further improved, *i.e.*, all metrics reach 0% on Protocol 1 of SiW.

Effect of the MIAD Module. From the results of CeFA dataset in Tab. V, we can see that the improvement of Aux.D $w/$ MIAD is more obvious than that of Aux.D $w/$ UPS when compared with Aux.D, *i.e.*, the value of ACER is reduced form 9.1% to 6.5% and 7.4% for the MIAD and UPS, respectively. It demonstrates that the MIAD module is better at detecting unseen types of attacks. Expected conclusions, using UPS and MIAD modules, that is, to optimize the backbone method from the input and feature level respectively, our approach improves the backbone performance and stability at the same time. We believe that other state-of-the-art methods can be further improved by equipping with our UPS and MIAD.

Furthermore, we calculate the mean and variance of the testing results for ten consecutive frames, and measure the performance and stability of the model, respectively. As shown in Fig. 3, the performance and stability of the Aux.D are 4.41% and 0.91% respectively. While our UPS module can greatly improve its performance and stability, *i.e.*, the mean ACER reduces to 1.58%, and the variance decreases to 0.28%. When the MIAD modules are further adopted, our method reaches the optimal values, and the mean and variance of ACER are 0.23% and 0.19% respectively.

F. Visualization Analysis

In this section, we visually analyze the effectiveness of improvement from our UPS module in detail. As shown the first and third rows in Fig. 4, we randomly select some sample pairs from OULU-NPU, SiW, and CeFA dataset. For each pair, it contains live face and fake face with different facial poses. After the first stage of using UPS module, the generated counterparts are shown in the second and fourth rows of Fig. 4, respectively. We can see that the UPS module generates

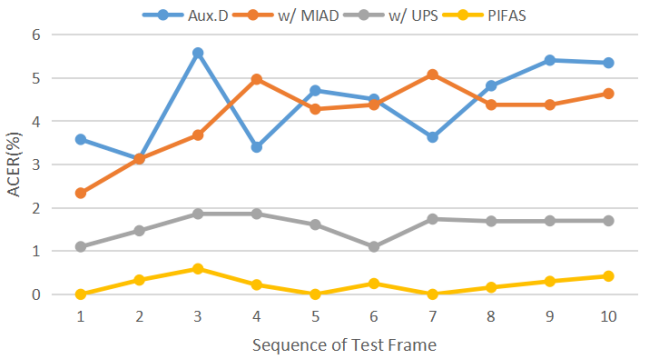

Fig. 3. Comparison of the 4 methods on the Protocol 1 of SiW dataset. It shows the results of ACER for ten consecutive frames of testing video.

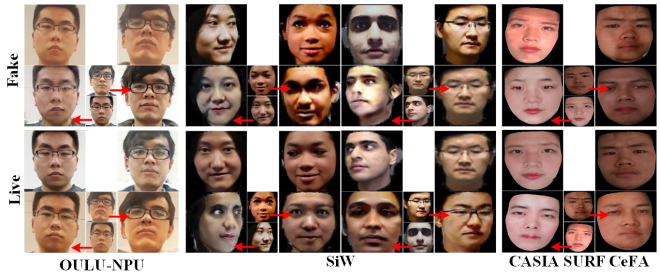

Fig. 4. Sample display on three datasets, where the first and third rows are fake and live faces respectively. The second and fourth rows are samples of the same category but different facial poses with the first and third rows, which the pose is specified by the sample in the lower right or upper left.

high-quality counterparts of the same category as the original samples but with different poses. Therefore, our UPS can alleviate the model's overfitting to some fixed facial poses by introducing samples with different poses at the input level.

IV. CONCLUSION

In this work, we propose a simple yet effective framework, namely PIFAS, against the inferences of liveness-irrelated factors. An UPS module is introduced as the first stage, which completes the facial pose replacement through a generative way. A MIAD module is adopted as the second stage, which incorporates knowledge about locality in the face into a score map. Extensive experiments demonstrate that the proposed method achieves competitive results on several benchmarks, especially for stabilizing the results of each testing video.

V. ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China 2021YFF0602103, the External cooperation key project of Chinese Academy Sciences 173211KYSB20200002, the Chinese National Natural Science Foundation Projects 61876179 and 61961160704, the Science and Technology Development Fund of Macau (0008/2019/A1, 0010/2019/AFJ, 0025/2019/AKP, 0004/2020/A1, 0070/2021/AMJ) and Guangdong Provincial Key R&D Programme: 2019B010148001.

REFERENCES

[1] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *ICB*, 2012.

[2] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *BIOSIG*, 2012.

[3] N. Erdogmus and S. Marcel, "Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect," in *BTAS*, 2014.

[4] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *CVPR*, 2019.

[5] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," *CVPR*, 2020.

[6] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma, "Face anti-spoofing via disentangled representation learning," in *ECCV*, 2020.

[7] Y. Liu, J. Stehouwer, and X. Liu, "On disentangling spoof trace for generic face anti-spoofing," in *ECCV*. Springer, 2020, pp. 406–422.

[8] A. Liu, Z. Tan, J. Wan, Y. Liang, Z. Lei, G. Guo, and S. Z. Li, "Face anti-spoofing via adversarial cross-modality translation," *TIFS*, vol. 16, pp. 2759–2772, 2021.

[9] Z. Chen, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, F. Huang, and X. Jin, "Generalizable representation learning for mixture domain face anti-spoofing," *arXiv preprint arXiv:2105.02453*, 2021.

[10] A. Liu, C. Zhao, Z. Yu, J. Wan, A. Su, X. Liu, Z. Tan, S. Escalera, J. Xing, Y. Liang *et al.*, "Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection," *arXiv preprint arXiv:2104.06148*, 2021.

[11] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *CVPR*, 2018.

[12] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *arXiv*, 2014.

[13] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *CCBR*, 2016.

[14] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *JVCIR*, 2016.

[15] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, and Z. Lei, "Deep spatial gradient and temporal depth learning for face anti-spoofing," *CVPR*, 2020.

[16] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *FG*, 2017.

[17] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *TIFS*, vol. 15, pp. 42–55, 2019.

[18] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *CVPR*, 2019, pp. 919–928.

[19] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante, and S. Z. Li, "Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing," *TBIOM*, vol. 2, no. 2, pp. 182–193, 2020.

[20] A. Liu, J. Wan, S. Escalera, H. Jair Escalante, Z. Tan, Q. Yuan, K. Wang, C. Lin, G. Guo, I. Guyon *et al.*, "Multi-modal face anti-spoofing attack detection challenge at cvpr2019," in *CVPRW*, 2019, pp. 0–0.

[21] A. Liu, C. Zhao, Z. Yu, A. Su, X. Liu, Z. Kong, J. Wan, S. Escalera, H. J. Escalante, Z. Lei *et al.*, "3d high-fidelity mask face presentation attack detection challenge," in *ICCVW*, 2021, pp. 814–823.

[22] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *NIPS*, vol. 29, pp. 2172–2180, 2016.

[23] T. Xiao, J. Hong, and J. Ma, "Elegant: Exchanging latent encodings with gan for transferring multiple face attributes," in *ECCV*, 2018.

[24] X. L. Amin Jourabloo*, Yaojie Liu*, "Face de-spoofing: Anti-spoofing via noise modeling," in *ECCV*, 2018.

[25] J. Stehouwer, A. Jourabloo, Y. Liu, and X. Liu, "Noise modeling, synthesis and classification for generic object anti-spoofing," in *CVPR*, June 2020.

[26] X. L. Yaojie Liu, Joel Stehouwer, "On disentangling spoof traces for generic face anti-spoofing," in *ECCV*, 2020.

[27] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018.

[28] B. Duan, C. Fu, Y. Li, X. Song, and R. He, "Cross-spectral face hallucination via disentangling independent factors," in *CVPR*, 2020.

[29] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in *IJCB*. IEEE, 2017, pp. 319–328.

[30] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *ICLR*, 2018.

[31] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *ECCV*, 2020.

[32] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017.

[33] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *WACV*, 2021, pp. 1179–1187.

[34] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu, "Face anti-spoofing: Model matters, so does data," in *CVPR*, 2019, pp. 3507–3516.

[35] Y. Qin, C. Zhao, X. Zhu, Z. Wang, Z. Yu, T. Fu, F. Zhou, J. Shi, and Z. Lei, "Learning meta model for zero- and few-shot face anti-spoofing," 2019.

[36] Z. Yu, Y. Qin, X. Li, Z. Wang, C. Zhao, Z. Lei, and G. Zhao, "Multi-modal face anti-spoofing based on central difference networks," in *CVPRW*, 2020.

[37] A. Parkin and O. Grinchuk, "Creating artificial modalities to solve rgb liveness," 2020.

[38] A. Liu, X. Li, J. Wan, Y. Liang, S. Escalera, H. J. Escalante, M. Madadi, Y. Jin, Z. Wu, X. Yu *et al.*, "Cross-ethnicity face anti-spoofing recognition challenge: A review," *IET Biometrics*, vol. 10, no. 1, pp. 24–43, 2021.