Uncertainty-Aware Physically-Guided Proxy Tasks for Unseen Domain Face Anti-spoofing

Junru Wu[1], Xiang Yu[2*], Buyu Liu[2], Zhangyang Wang[3] and Manmohan Chandraker[2]

[1]Department of Computer Science and Engineering, Texas A&M University, 435 Nagle Street, College Station, 77843, TX, USA.
[2]NEC Laboratories America, Inc., 2033 Gateway Place Suite 200, San Jose, 95110, CA, USA.
[3]Department of Electical and Computer Engineering, The University of Texas at Austin, 2501 Speedway, Austin, 78712, TX, USA.

*Corresponding author(s). E-mail(s): xiangyu@nec-labs.com;
Contributing authors: sandboxmaster@tamu.edu; buyu@nec-labs.com; atlaswang@utexas.edu; manu@nec-labs.com;

Abstract

Face anti-spoofing (FAS) seeks to discriminate genuine faces from fake ones arising from any type of spoofing attack. Due to the wide varieties of attacks, it is implausible to obtain training data that spans all attack types. We propose to leverage physical cues to attain better generalization on unseen domains. As a specific demonstration, we use physically guided proxy cues such as depth, reflection, and material to complement our main anti-spoofing (a.k.a liveness detection) task, with the intuition that genuine faces across domains have consistent face-like geometry, minimal reflection, and skin material. We introduce a novel uncertainty-aware attention scheme that independently learns to weigh the relative contributions of the main and proxy tasks, preventing the over-confident issue with traditional attention modules. Further, we propose attribute-assisted hard negative mining to disentangle liveness-irrelevant features with liveness features during learning. We evaluate extensively on public benchmarks with intra-dataset and inter-dataset protocols. Our method achieves the superior performance especially in unseen domain generalization for FAS.

Keywords: Face-Anti Spoofing, Physically-Guided, Uncertainty-Aware

I Introduction

With growing prevalence of face recognition, it is increasingly subject to a wide variety of spoofing attacks. Thus, face anti-spoofing or liveness detection is emerging as an essential precursor, with the key need being robust to various attacks that are possibly unseen previously and drastically different in appearance from training data. This is an extremely challenging problem due to the fact that sophisticated spoofs might arise from similar camera and lighting setups as genuine inputs, leading to only subtle differences in appearance. On the other hand, types of attacks range from printed photos to facial masks, which makes it laborious to obtain exhaustive training data for anti-spoofing task.

In this paper, we aim to address the face anti-spoofing problem on *unseen* domains or attack types, where neither definition of the attack types nor training data under supervised or unsupervised condition is available. To achieve this, we derive inspiration from physical cues that establish a commonality for

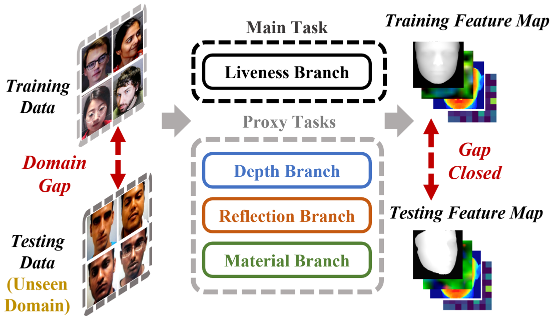


Fig. 1: Traditional methods only focus on real/spoofing binary classification, which results in sensitive prediction. By introducing the physical cues of depth, material and reflection as proxy tasks, our method largely close the domain gap from the training data to the unseen testing data and thus boost the performance reliably.

genuine inputs and distinction from fake ones. We refer to the estimation of these cues as *proxy tasks*, performed in conjunction with the main task, a.k.a appearance-based liveness detection. While the proposed formulation is general and physical cues can be arbitrary, we focus on depth estimation, reflection detection and material classification as our proxy tasks. Intuitively, we expect genuine faces to constitute face-like geometry and present skin as the material, while several presentation attacks might violate at least one of those conditions. As a consequence, incorporating such proxy tasks enables to generalize the shared cues to unseen domains or attack types.

We bring the insights from single-image based face reconstruction using 3D morphable models (3DMM) [1] for depth proxy, single-image based material recognition trained on large-scale datasets [2] for material proxy, and a single image reflection separation model [3] to provide the pseudo labels for the reflection proxy. A shared encoder is trained across the main and proxy tasks to transfer the insights into our deep appearance-based liveness detection problem. In contrast to existing works [4] that incorporates depth cue to regularize sensitive binary liveness task, our proxy tasks are more general in the sense of considering more physical cues such as material, to gear towards a physically meaningful way to handle unseen domains. Meanwhile, we organize the proxy tasks into a multi-channel learning framework to provide a more robust detection with an attention aggregation. Note that domain adaptation is not applicable in our setting, since we assume that even unlabeled training data is not available, which cannot define the target domain.

Besides proxy tasks, we also leverage a *pretext task* in the form of face recognition, which is usually regularized by large scale labeled datasets and expected to provide high-level shared face analysis feature representation for liveness detection. We thereby provide recipes for pre-training on face recognition that allows better generalization to unseen domains for the liveness task, as well as multi-channel training with liveness and proxy tasks. We conduct extensive experiments on five publicly available benchmarks. In each case, we demonstrate not only state-of-the-art results, but also that judicious use of pretext and proxy tasks allows better generalization of liveness detection to unseen domains. Besides, the multi-task learning can result in channel conflict as the liveness feature is ideally invariant to identity information where the pretext task and our liveness task share the same network. To this end, we leverage the attribute information in the proxy data to conduct a triplet metric learning based mining, expecting to better disentangle the non-liveness information from the learned feature and thus boosts the liveness detection.

To better exploit multiple physically meaningful resources, we further holistically weigh the relative contributions of the main task and various proxy tasks with an uncertainty-aware attention module. Traditional attention modules are jointly optimized with all tasks and might cause the notorious over-confident issue due to training data bias. While our uncertainty-aware attention is designed to independently estimate the tasks' variance, which does not capture feature fitness to the task but rather focusing on its deviation to the estimated mean or termed uncertainty of the feature estimation. This property ensures less bias in uncertainty-aware attention module thus captures the property of input images better.

In summary, we propose the following contributions:

- We propose three physical-cue guided proxy tasks including depth, material and reflection, which share the commonality across domains to enable the unseen domain anti-spoofing.
- We leverage an uncertainty-aware attention module to effectively combine the main and proxy tasks and boost the performance.
- We propose an attribute-assisted mining scheme to make sure liveness-irrelevant features are properly disentangled and only liveness features are learned.

Method	Training								Evaluation
	Physical Cues						Temporal Cues	Multi-domain Cues	Cross-domain Eval
	Blink	Depth	rPPG	Reflection	Material	Point Cloud			
[5]								✓	✓
[6]	✓						✓		✓
[4]		✓	✓						✓
[7, 8]		✓						✓	
[9]				✓					
[10]		✓		✓		✓			
[11]		✓		✓					
[12]		✓		✓	✓				
Ours		✓		✓	✓				✓

Table I: Comparison with other unseen domain anti-spoofing methods, "✓" denotes applicable.

- We conduct an extensive evaluation with both intra-dataset and inter-dataset protocols including the latest attribute-rich CelebA-Spoof dataset, highlighting our framework's better performance in unseen domain generalization for FAS.

II Related Work

We categorize face anti-spoofing literature into physical cue based, feature learning based methods, and whether they address the unseen spoofing attacks. A overall method comparison is listed in Table I.

Physical Cue based Anti-spoofing: Early research on anti-spoofing leverages the physical cues, e.g. head movement [13] and eye blinking [14], to indicate the genuineness. These methods can be simply spoofed by printing faces with eye region cut, or wearing a facial mask and moving head. By analyzing the lighting cooperated from different reflection, the remote Photoplethysmography (r-PPG) [15–18] is proposed to identify spoofing attacks with material information. However, this type of methods require the imaging quality to be high as the lighting measurement is less tolerated to noise. Combining with CNN, depth is proposed [4, 7] as an auxiliary task that enables less sensitive and more explainable training. [19] introduced reflection map as a supplement to depth for bipartite auxiliary supervision. This method is limited in generalizing to other spoofing resources since their depth, reflection and r-PPG are trained on a single dataset. Instead, our depth, reflection and material channels are guided by models trained on large scale datasets, i.e., a 3DMM based depth regression model [20], a single image reflection separation model [3] and a material classification model [2], which substantially improves the generalization.

Learning based Anti-spoofing: The handcrafted features, e.g., HoG [21, 22], SIFT [23] and LBP [24–28] are explored in early literature. Such binary classification achieves good performance but is restricted to some defined domains. Meanwhile, those methods do not consider environment variations, i.e., lighting, color tone or pose change. To this end, the HSV and YCbCr [24], Fourier transform [29] and image low-rank decomposition [30] methods are also explored. Some other works [31–35] utilize the temporal information assuming videos are available. Later, deep learning based features [4, 6, 34, 36–39] are utilized [4, 6, 34, 36–39] and achieve better performance. Notice that both [38] and [4] leverage texture and depth, which seem to be close to our setting. However, instead of directly exploring texture, we formulate the texture into a more physically consistent cue, the material, and set up the material classification task to avoid rPPG calculation. Moreover, our method is single image based which does not require the temporal information, thus reducing the run-time and model complexity. DRL-FAS[9] proposed a Destruction and Combination Module (DCN) to learn a more robust feature via permuting patches, while using a reflection estimation network as auxiliary task. [8] use a contrastive Depth Loss for more accurate depth supervision. [10] use pixel-wise dense supervision such as Binary Mask, Reflection, Depth and 3D Point Cloud prediction to improve the performance and enhance the model's interpretability. [12] utilize discrepancy across different physical materials to extract discriminative and robust features for FAS.

To learn better representations, [40] proposed to use Reinforcement Learning with a recurrent mechanism to learn local information sequentially from the explored domain. [41] use a multi-branch approach to decomposite the high-frequency and low-frequency information. [42] propose a Cross-modal Auxiliary (CMA) framework to close the visible gap between different modalities via mapping inputs from one modality to another, [43] extract shading-based 3D features from a pair of images captured under different illumination. [44] introduce Central Difference Convolution (CDC) into antispoofing, while [45][46] further incorporate Neural Architecture Search (NAS) to build a more powerful and efficient network structure.

Unseen Domain Anti-spoofing: Methods that explore handcrafted features can deal with unseen spoofing attacks as these features are independent from attack types. However, due to the limitation of feature representation power, they cannot generalize well. Patel et al. [6] propose to combine deep features with eye blinking cues for cross-dataset spoofing detection. [47] use a semi-supervised learning framework with a adaptive transfer mechanism to alleviate the influence of unseen spoofing data. [48] adaptively selects feature normalization methods according to the inputs to improve generalization performance. [49] use a adaptive sampling strategy that iteratively reweights the sample importance to further improve the generalization. Other works [50, 51] formulating the anti-spoofing task into an anormaly or outlier detection, which highly rely on the definition of genuine samples. There are also works achieve the goal of domain generalization via meta-learning, for example, [52] propose domain dynamic adjustment meta-learning that iteratively divides mixture domains, [53] define FAS as a zero- and few-shot learning problem and tackle it through meta-learning. [54] train a meta teacher in a bi-level optimization manner to supervise FAS detector effectively. To alleviate this, Liu et al. [5] propose a zero-shot learning solution, whereas the unseen attacks are assigned to the most similar attacks predefined in the database. These unseen attacks are wildly variant and leave the chance that the attacks are heavy outliers. Shao et al. [7] formulate a domain generalization approach to improve the generalization ability, which depends on the number and diversity of the seen domains, i.e., biased or long-tailed observed domains would degrade the performance. Different from [5, 7], we propose the physical cue based proxy tasks that are less dependent on the data distribution, which generally could be more stable and consistent across seen and unseen attacks.

Uncertainty Analysis: Uncertainty provides an effective measurement for model/data reliability [55–58]. It has been widely applied in many vision tasks such as classification [59], semantic segmentation [60] and face recognition [61]. Our method follows the setting of [60] by leveraging multiple tasks, but in a complete different problem as face anti-spoofing rather than segmentation. We consider our proxy tasks are orthogonal to each other, whereas in [60] those semantic tasks are strongly correlated. To the best of our knowledge, we are the first to leverage uncertainty in face anti-spoofing tasks.

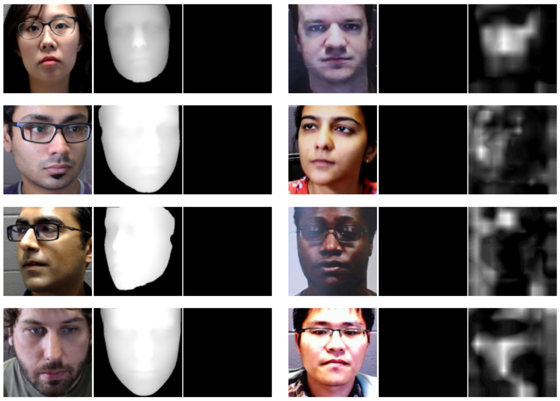

(a) Genuine faces (b) Spoof faces

Fig. 2: (a) From Left to Right: examples of face image, depth map, reflection map for genuine faces (b) From Left to Right: examples of face image, depth map, reflection map for spoof faces.

III Proposed Approach

In this section, we firstly introduce the shared feature extractor learning by incorporating the pretext task face recognition. Then, the physical cue based proxy tasks, i.e., depth estimation, reflection detection and material classification, are introduced as the spoofing attack detection anchors. Finally, an uncertainty-aware attention module is proposed to aggregate the proxy channels for optimal performance.

III.A Shared Feature Representation Learning

As shown in Figure 3, our framework consists of multiple channels of pretext and proxy tasks. Separating each single task with independent CNNs results in network redundancy. Moreover, the separated CNNs cannot leverage the rich information from the other tasks, where hyper-column [62] and deeply-supervised net [63] have shown a highly integrated

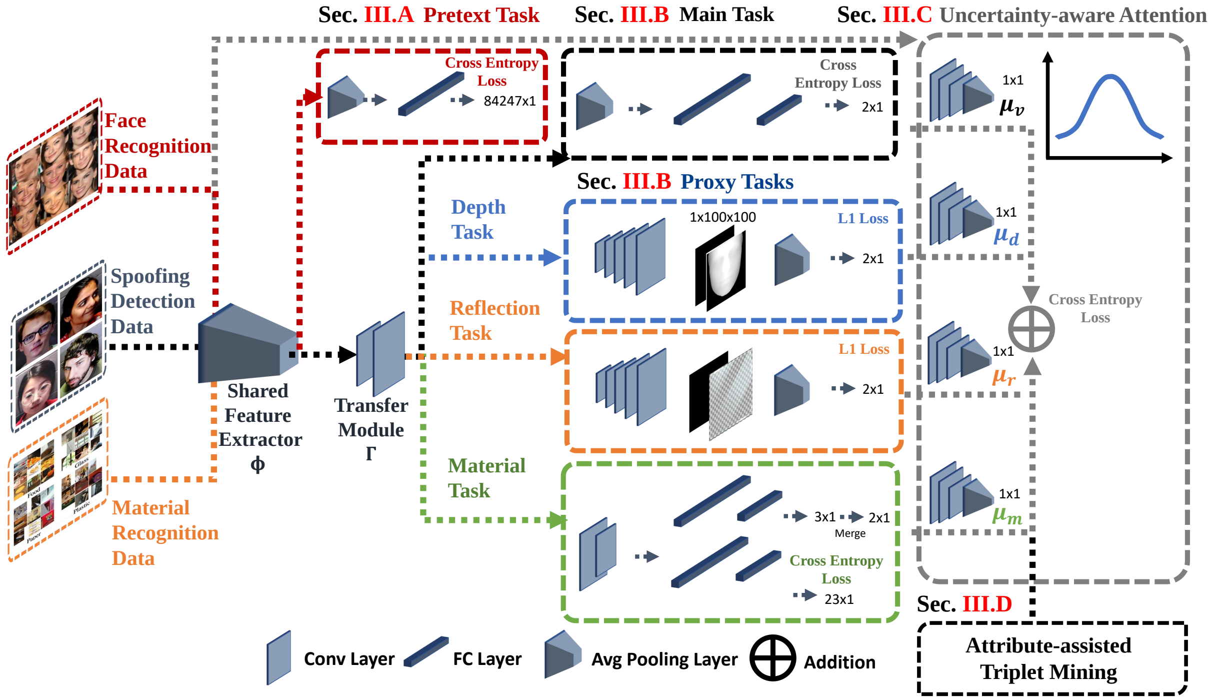


Fig. 3: The proposed framework consists of the pretext task "face recognition" (Sec. III.A), the proxy tasks (Sec. III.B) "depth estimation","material prediction", "reflection detection" and the main task "liveness detection". A novel triplet mining regularization (Sec. III.D) is proposed to better disentangle the liveness feature and an uncertainty-aware attention module (Sec. III.C) aggregates the channel-wise results for boosted performance.

framework for multiple tasks is beneficial. To this end, we propose to use a single feature extractor Φ to provide the shared feature for all the downstream tasks.

The shared features should provide high-level task-specific yet general information for downstream tasks such that we neither drift away from original tasks nor learn only task-driven representations. Among the face analysis applications, face recognition is a promising pretext as it is usually trained with large-scale data including millions of identities, which guarantees the robustness as well as the discriminality. Other candidates such as facial attribute classification, expression recognition or spoofing detection are not general or robust, as each of the tasks conduct a 10-way or 2-way classification, which can be sensitive or easily overfitting [4]. Thus, to initialize the feature extractor Φ, we apply face recognition as our pretext task.

Denoting input image as $\mathbf{x}_r, \mathbf{x}_v$ and $\mathbf{x}_m$ for recognition, spoofing and material data respectively. After the shared feature extractor Φ, the pretext task applies a filter Ψ_r to refine the face identity feature. The loss is:

$$\mathcal{L}_r = -\sum_i \mathbb{1}(y_i) \log \frac{\exp(\mathbf{w}_i \Psi_r \circ \Phi(\mathbf{x}_r))}{\sum_j \exp(\mathbf{w}_j \Psi_r \circ \Phi(\mathbf{x}_r))} \quad (1)$$

where y_i is the ground truth label for identity i. j varies across the whole number of identities. $\mathbf{w}_i$ is the i_{th} separation hyper-plane of the classifier. $\circ$ denotes the sequential network flow.

III.B Multi-channel Proxy Task Learning

We introduce a transfer module Γ to adapt the rich feature extracted by Φ into the spoofing detection related tasks. Directly utilizing the shared feature leads to sub-optimal prediction as it incorporates unrelated face recognition cues, which may serve as noise. Similar to the pretext task, we set up multiple channels for our proxy tasks, i.e., liveness detection Ψ_v, depth estimation Ψ_d, reflection detection Ψ_r and material prediction Ψ_m.

Liveness Detection Main Task: The spoofing detection is a well-known binary classification task. The input is spoofing face $\mathbf{x}_v$. After shared extractor and feature transfer module, we set up the spoofing detection channel filter Ψ_v to conduct the binary classification task, in which we adopt binary cross entropy loss as the objective:

$$\mathcal{L}_v = -y_v \log(\mathbf{p}(\mathbf{z})) - (1 - y_v)\log(1 - \mathbf{p}(\mathbf{z})) \quad (2)$$

$$\mathbf{p}(\mathbf{z}) = \frac{\exp(\tilde{\mathbf{w}}_0\mathbf{z})}{\exp(\tilde{\mathbf{w}}_0\mathbf{z}) + \exp(\tilde{\mathbf{w}}_1\mathbf{z})} \quad (3)$$

where $\mathbf{y}_v$ is the ground truth of spoofing or genuine, $\mathbf{z} = \Psi_v \circ \Phi(\mathbf{x}_v)$ denotes the spoofing detection feature after the spoofing detection filter Ψ_v. $\tilde{\mathbf{w}}_0$ and $\tilde{\mathbf{w}}_1$ are the separation hyper-planes of the binary classifier. Likelihood of being spoofing sample $\mathbf{p}(\mathbf{z})$ is estimated via a softmax operation in Equation 3.

Depth Proxy Task: We believe the physical cues should share similar characteristics for genuine faces across different attack types or spoofing datasets. Thus the depth prediction should also be consistent. We aim to predict the per-pixel depth map given the input face image. We leverage an hourglass network structure to conduct this regression problem, which has been proved effective in key point detection [64] and image segmentation [65]. To prepare the ground truth depth map $\mathbf{d}_{GT}$, we apply a 3D face shape reconstruction algorithm [20] offline to estimate the dense point cloud for the face images. As for genuine face image, we utilize the predicted depth as its ground truth depth map, where background is set as 0. For 2D spoofing face images, according to their attack types, i.e., display screen or paper, we know that the actual depth is from a flat plane of either screen or paper. Thus, we manually set the spoofing ground truth depth to be all 0. The absolute depth is unnecessary since we only focus on the relative face geometry. We show some examples of generated depth map results in Figure 2. During training, an l_1-based reconstruction loss is applied:

$$\mathcal{L}_d = \|\Psi_d \circ \Gamma \circ \Phi(\mathbf{x}_v) - \mathbf{d}_{GT}\|_1 \quad (4)$$

where Ψ_d is the hourglass net depth estimation module, Γ is the feature transfer module and $\mathbf{d}_{GT}$ is the ground truth depth. Notice that for depth estimation, we input the spoofing data $\mathbf{x}_v$ with the augmented ground truth depth map. We do not utilize extra depth data for this channel.

Reflection Proxy Task: Reflection is another useful physical cue that indicates the genuine faces, as non-skin materials inevitably show abnormal reflection compared to skin. As a result, for spoofing face, we use a single image reflection separation model [3] to generate the reflection map, while for genuine face, we set the it to zero denoting no reflection is present. Visual examples of generated reflection maps are in Figure 2. During training, we push the predicted reflection map to be close to the pseudo ground truth under l_1-based constraint:

$$\mathcal{L}_r = \|\Psi_r \circ \Gamma \circ \Phi(\mathbf{x}_v) - \mathbf{r}_{GT}\|_1 \quad (5)$$

where Ψ_r is the hourglass net reflection estimation moduleand $\mathbf{r}_{GT}$ is the ground truth reflection map.

Material Proxy Task: Though reflection in some way indicates the material information, we explicitly introduce material as another proxy task to leverage the correlation among the multiple tasks, expecting to benefit from the multi-task learning. The physical insight for material in liveness detection is that skin across different spoofing attacks or spoofing datasets should remain similar RGB information. We automatically obtain the material type for face spoofing data according to its attack type. For instance, we denote the material class of screen display and paper print as "glass" and "paper" respectively. In this way, we actually unify the material type towards the general material recognition [2].

Notice that the number of material types in spoofing data can be limited, which may encounter the same sensitivity issue as the binary spoofing detection task. To this end, we introduce the general material recognition data to anchor the material feature space from being collapsed. Specifically, the general material recognition and our spoofing data material recognition share all the network structures except the last classifier layer. As in general material recognition, there are 23 defined categories [2], such as brick, metal, plastic, skin, glass, etc. We set up a 23-way classifier $\mathbf{C}_g$ for the general material recognition and a 3-way classifier $\mathbf{C}_v$ for our spoofing data material recognition. A multi-source scheme is proposed to train the modules of Φ, Γ and Ψ_m jointly. Denoting the feature $\mathbf{f} = \Psi_m \circ \Gamma \circ \Phi(\mathbf{x})$, a combined multi-class softmax loss is applied to train $\mathbf{C}_g$ and $\mathbf{C}_v$:

$$\mathcal{L}_m = -\sum_{i=1}^{23} \mathbb{1}(l_i) \log \frac{\exp(\omega_i \Psi_m \circ \Gamma \circ \Phi(\mathbf{x}_m))}{\sum_j \exp(\omega_j \Psi_m \circ \Gamma \circ \Phi(\mathbf{x}_m))} -$$

$$\sum_{i=1}^{3} \mathbb{1}(l_i) \log \frac{\exp(\tilde{\omega}_i \Psi_m \circ \Gamma \circ \Phi(\mathbf{x}_v))}{\sum_j \exp(\tilde{\omega}_j \Psi_m \circ \Gamma \circ \Phi(\mathbf{x}_v))} \tag{6}$$

where l_i is the material ground truth label, ω_i and ω_j, $\tilde{\omega}_i$ and $\tilde{\omega}_j$ are the separation hyper-planes for $\mathbf{C}_g$ and $\mathbf{C}_v$ respectively. By alternatively feeding the material and spoofing data, we guarantee that $\mathbf{m}$ is generalized for not only the standard material recognition, but also the material recognition for face spoofing data.

III.C Uncertainty-aware Attention Modeling

As each of the channels looks into different aspects of the spoofing characteristics, we seek to combine those independent channels adaptively to boost the final spoofing detection performance. We introduce an uncertainty-driven attention module that is orthogonal to each of the main and proxy tasks, which thus effectively overcomes the over-confident issue of the traditional attention modules.

Given an input $\mathbf{x}_v$, the joint likelihood $p(y \mid \mathbf{x}_v) = p(\mathbf{z} \mid \mathbf{x}_v)p(\mathbf{d} \mid \mathbf{x}_v)p(\mathbf{r} \mid \mathbf{x}_v)p(\mathbf{m} \mid \mathbf{x}_v)$ according to the channel independence assumption, where $\mathbf{z}$ is from Equation 3 as the main task feature, $\mathbf{d} = \Psi_d \circ \Gamma \circ \Phi(\mathbf{x}_v)$ is from Equation 4 as the reflection feature, $\mathbf{r} = \Psi_r \circ \Gamma \circ \Phi(\mathbf{x}_v)$ is from Equation 5 as the depth feature, $\mathbf{m} = \Psi_m \circ \Gamma \circ \Phi(\mathbf{x})$ is from Equation 6 as the material feature. Maximizing the joint likelihood leads to maximizing the summation of each likelihood:

$$\arg\min - \log(p(y \mid \mathbf{x}_v) = - \sum_{\mathbf{u}=\mathbf{z},\mathbf{d},\mathbf{r},\mathbf{m}} \log(p(\mathbf{u} \mid \mathbf{x}_v) \tag{7}$$

For each channel, we assume a Gaussian distribution $p(\mathbf{u} \mid \mathbf{x}_v) \sim \mathcal{N}(\mu_{\mathbf{u}}, \sigma_{\mathbf{u}})$ to capture the uncertainty. For liveness/material classification task $\mu_{\mathbf{u}}$ is the separation hyper-plane vector, while for the depth/reflection regression task $\mu_{\mathbf{u}}$ is the mean prediction map on the training set. Under the probabilistic setting, such $\mu_{\mathbf{u}}$ conforms to another Gaussian distribution $\mathcal{N}(\mu_{\mathbf{u}}, \sigma_{\mu_{\mathbf{u}}})$, where $\sigma_{\mu_{\mathbf{u}}}$ is estimated upon sampling from multiple rounds of training. $\mu_{\mathbf{u}}$ is independently learned via a 3-layer ConvNet conditioned on the input $\mathbf{x_v}$. $\mu_{\mathbf{u}}$ is jointly learned with $\mathbf{u}$ during training and is fixed during inference. The objective to learn $\sigma_{\mathbf{u}}$ is defined in the following:

$$\mathcal{L}_{\sigma_{\mathbf{u}}} = \sum_{\mathbf{u}=\mathbf{z},\mathbf{d},\mathbf{r},\mathbf{f}} \left(\frac{\|\mathbf{u} - \mu_{\mathbf{u}}\|^2}{2(\sigma_{\mathbf{u}}^2 + \sigma_{\mu_{\mathbf{u}}}^2)} + \frac{D}{2} \log(\sigma_{\mathbf{u}}^2 + \sigma_{\mu_{\mathbf{u}}}^2) \right) \tag{8}$$

where D is the feature dimension. It is independently optimized after the network is converged. During inference, the network outputs $\mu_{\mathbf{u}}$ and $\sigma_{\mu_{\mathbf{u}}}$ simultaneously. Given $\mu_{\mathbf{u}}$ and $\sigma_{\mu_{\mathbf{u}}}$, we then obtain the uncertainty estimate for each channel with Equation 8.

To sum up, we propose a two-stage training procedure. The first stage consists of the training of liveness main task, proxy tasks and pretext task. And the loss is defined as following:

$$\mathcal{L} = \lambda_v \mathcal{L}_v + \lambda_d \mathcal{L}_d + \lambda_r \mathcal{L}_r + \lambda_m \mathcal{L}_m + \lambda_t (\mathcal{L}_{tid} + \mathcal{L}_{ta}) \tag{9}$$

Then in the second stage, the uncertainty attention module is trained with Equation 8.

III.D Attribute-assisted Triplet Mining

To better disentangle the liveness feature apart from identity information and other facial attributes information, we leverage the metric learning to regularize the feature representation learning. Specifically, given the input $\mathbf{x}_v^i$, we would expect the following loss to be minimized such that the identity information can be decoupled from the liveness feature.

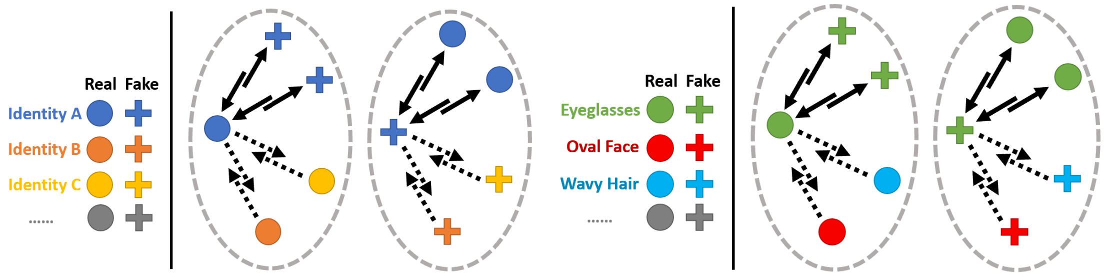
$$\mathcal{L}_{tid} = \lfloor \|\Phi(x_v^{i,j}) - \Phi(x_v^{i,k})\|^2 - \|\Phi(x_v^{i,j}) - \Phi(x_v^h)\|^2 + m_1 \rfloor_+ \tag{10}$$

$x_v^{i,j}$ means the j^{th} liveness sample from identity i, while x_v^h simply means the liveness sample from other identities as a negative sample.

Similarly for other face attributes introduced in CelebA [11, 66], we believe the orthogonality can be preserved if those attribute information is disentangled from the liveness information.

$$\mathcal{L}_{ta} = \lfloor \|\Phi(x_v^{a_i,j}) - \Phi(x_v^{a_i,k})\|^2 - \|\Phi(x_v^{a_i,j}) - \Phi(x_v^{a_h})\|^2 + m_2 \rfloor_+ \tag{11}$$

a_i indicates an attribute label and a_h indicates a different attribute label for the negative sample. m_1 and m_2



(a) Disentangling **Identity attributes**

(b) Disentangling **Face attributes**

Fig. 4: The illustration of proposed Attribute-assisted Triplet Mining. Here we give two examples (a) disentangling identity on SiW[4] dataset (b) disentangling face attributes on CelebA-Spoof[11] dataset, respectively.

here are the margin hyper-parameter set to squeeze the classification boundary for better feature learning.

IV Implementation Details

In our implementation, we leverage a pre-trained face recognition engine and re-utilize the encoder as our shared feature extractor Φ. Then, we keep the face recognition as our pretext task and equip the main and proxy tasks to form a multi-source multi-channel training. As illustrated in the methodology section, a two-stage training is conducted. For the first stage joint training of pretext, main and proxy tasks, we apply random cropping and horizontal flipping as data augmentation.

We adopt Adam solver and the initial learning rate is set 0.0001. The momentum and weight decay are fixed as 0.9 and 0, respectively. Hyper-parameters in Equation 9 is empirically searched via some hold-out validation as $\lambda_v = 1, \lambda_d = 0.1, \lambda_r = 0.1, \lambda_m = 0.1$, $\lambda_t = 0.1$ for triplet need to add here together with m_1 and m_2 as in Equation 10 and 11 . respectively. For the second stage, when training the uncertainty-aware attention module, we re-use well-trained modules from the first stage, and only fine-tune the two-layer fully connected layers for each of the main and proxy tasks to estimate the variance.

V Experiments

V.A Datasets

CASIA [68]: A video based 2D spoofing attack database, consists of 600 videos from 50 people, 240 videos from 20 people for training and 360 videos from 30 people for testing. Each people contains 12 videos with video re-display and photo print attacks, of which 8 are normal resolution videos and 4 are high resolution videos. The photo attacks are further categorized into cut photo by cutting holes around eyes, noise, mouth, and warped photos by warping photos with different curvature.

Replay-Attack [69]: A 2D face spoofing attack database consists of $1,300$ video clips of photo and video attack attempts from 50 clients, under different lighting conditions. To produce the attacks, high-resolution photos and videos from each client were taken under the same conditions as in their authentication step.

MSU-MFSD [70]: it consists of 280 video clips of photo and video attack attempts from 35 clients. Mobile phones are used to capture both genuine faces and spoofing attacks. Printed photos are generated from high quality color printers for another attack type. It also provides replay video attacks with high resolution of 2048×1536 from iPad air screen.

Oulu-NPU [67]: A large-scale 2D spoofing attack dataset, consists of 4950 genuine and attack videos from 55 people. They are recorded using the front cameras of different mobile devices with variant illuminations and backgrounds. The attack types are print and video replay. There are four protocols designed to consider generalization on cross attack types and capturing sensor types.

SiW [4]: Spoofing in the Wild dataset provides 4,478 genuine and spoofing videos from 165 subjects. For each subject, 8 genuine and up to 20 spoofing videos are captured. It systematically considers variations from subjects, camera sensors, spoofing attack types, lighting conditions, image resolution, and different

	Proxy Tasks			Components		Protocol 1			Protocol 2			Protocol 3		
Metrics	D	M	R	Att.	Tri.	APCER↓	BPCER↓	ACER↓	APCER↓	BPCER↓	ACER↓	APCER↓	BPCER↓	ACER↓
SVM$_{RBF}$+LBP[67]						4.17	4.17	4.17	5.29±4.39	5.29±4.39	5.29±4.39	16.84±1.89	16.84±1.89	16.84±1.89
SVM$_{RBF}$+BSIF [50]						7.95	7.95	7.95	7.34±3.30	7.34±3.30	7.34±3.30	25.56±5.63	25.56±5.63	25.56±5.63
FAS-BAS[4]						3.58	3.58	3.58	0.57±0.69	0.57±0.69	0.57±0.69	8.31±3.81	8.31±3.81	8.31±3.81
FAS-TD-SF[27]						1.27	0.83	1.05	0.33±0.27	0.29±0.39	0.31±0.28	7.70±3.88	7.76±4.09	7.73±3.99
Ours						0.68	0.68	0.68	0.42±0.37	0.42±0.37	0.42±0.37	7.97±5.03	7.97±5.03	7.97±5.03
Ours (+Proxy Tasks)	✓					0.47	0.47	0.47	0.25±0.21	0.25±0.21	0.25±0.21	7.75±4.97	7.75±4.97	7.75±4.97
	✓	✓				0.57	0.57	0.57	0.27±0.24	0.27±0.24	0.27±0.24	7.80±4.95	7.80±4.95	7.80±4.95
		✓	✓			0.62	0.62	0.62	0.34±0.30	0.34±0.30	0.34±0.30	7.92±5.01	7.92±5.01	7.92±5.01
	✓		✓			**0.42**	**0.42**	**0.42**	0.27±0.22	0.27±0.22	0.27±0.22	7.73±4.96	7.73±4.96	7.73±4.96
	✓	✓	✓			0.44	0.44	0.44	**0.24±0.22**	**0.24±0.22**	**0.24±0.22**	**7.52±4.91**	**7.52±4.91**	**7.52±4.91**
Ours (+Attention)	✓			✓		0.46	0.46	0.46	0.25±0.22	0.25±0.22	0.25±0.22	7.50±4.79	7.50±4.79	7.50±4.79
	✓	✓		✓		0.55	0.55	0.55	0.26±0.22	0.26±0.22	0.26±0.22	7.76±4.90	7.76±4.90	7.76±4.90
		✓	✓	✓		0.60	0.60	0.60	0.31±0.29	0.31±0.29	0.31±0.29	7.88±4.96	7.88±4.96	7.88±4.96
	✓		✓	✓		0.41	0.41	0.41	0.23±0.21	0.23±0.21	0.23±0.21	7.43±4.82	7.43±4.82	7.43±4.82
	✓	✓	✓	✓		**0.39**	**0.39**	**0.39**	**0.23±0.20**	**0.23±0.20**	**0.23±0.20**	**7.39±4.72**	**7.39±4.72**	**7.39±4.72**
Ours (+Triplet Mining)	✓	✓	✓	✓	✓	**0.36**	**0.36**	**0.36**	**0.20±0.16**	**0.20±0.16**	**0.20±0.16**	**7.32±4.80**	**7.32±4.80**	**7.32±4.80**

Table II: Ablation Study of the Intra-dataset evaluation on SiW dataset. **D**,**M**,**R** denote depth, material, reflection proxy tasks, respectively. **Att.** denote Uncertainty-aware Attention Modeling. **Tri.** denote Attribute-assisted Triplet Mining. ACER and AUC (%) is reported. Best results are shown in **Bold**.

Attack Type	Proxy Tasks			Video		Digital Photo		Printed Photo	
Methods	D	M	R	ACER↓	AUC(%)↑	ACER↓	AUC(%)↑	ACER↓	AUC(%)↑
NN+LBP[51]				–	99.75	–	95.17	–	78.86
GMM+LBP[51]				–	93.20	–	87.80	–	89.19
OC-SVM$_{RBF}$[50]				–	91.55	–	84.97	–	87.19
SVM$_{RBF}$+LBP[67]				2.69	–	3.53	–	1.95	–
SVM$_{RBF}$+BSIF[50]				6.17	–	1.27	–	12.03	–
DTL[5]				–	99.90	–	99.90	–	99.60
Ours				0.20	99.93	1.80	99.80	1.71	99.60
	✓			0.10	99.99	0.37	99.99	0.88	99.99
	✓	✓		0.09	99.99	0.37	99.99	0.87	99.99
	✓	✓	✓	**0.09**	**99.99**	**0.32**	**99.99**	**0.82**	**99.99**

Table III: Ablation Study of the Intra-dataset evaluation on Replay-Attack dataset across Video, Digital Photo and Printed Photo. Best results are shown in **Bold**. ACER and AUC (%) is reported.

Methods	Training Set	Proxy Tasks			Extra Tasks		Compon.		HTER(%)↓
		D	M	R	I	S	Att.	Tri.	
FAS-TD-SF [27]	SiW								39.4
AENet[11]	CelebA-Spoof								14.3
AENet$_{C,G}$ [11]		✓		✓					14.1
AENet$_{C,S}$ [11]					✓	✓			12.1
AENet$_{C,S,G}$ [11]		✓		✓	✓	✓			11.9
Ours	CelebA-Spoof								14.6
		✓					✓		12.8
		✓		✓			✓		11.6
		✓	✓				✓		12.0
		✓	✓	✓			✓		11.3
		✓	✓	✓			✓	✓	**10.1**

Table IV: Inter-dataset benchmark results with CelebA-Spoof dataset, all the models are tested on CASIA dataset. Note that **I**,**S** denote Illumation, Spoof Type tasks introduced in [11], respectively. HTER (%) is reported.

sessions for capturing. There are three evaluation protocols emphasizing the generalization on face PAD, cross attack types, and unknown attack types.

CelebA-Spoof [11]: CelebA-Spoof is the current largest face anti-spoofing dataset containing 625,537 images from 10,177 subjects. The dataset featured rich annotations, which includes 43 rich face attributes, illumination, environment and spoof types. The live image is inherented from the CelebA dataset, while the spoof images is newly collected in controlled environment. Among 43 rich attributes, 40 attributes belong to Live images including all facial components and accessories such as skin, nose, eyes, eyebrows, lip, hair, hat, eyeglass. 3 attributes belong to spoof images including spoof types, environments and illumination conditions.

V.B Evaluation Metrics

The evaluation is focused on testing the generalization of cross attack types within one dataset, termed intra-dataset evaluation, and cross dataset spoofing detection, termed inter-dataset evaluation following [50]. To be consistent with most of the previous spoofing detection works, we apply the evaluation metrics as: Attack Presentation Classification Error Rate (APCER[71]), Bona Fide Presentation Classification Error Rate (BPCER[71]), ACER = 0.5(APCER+BPCER), Area Under Curve (AUC) ratio and Half-Total Error Rate (HTER). Further following [4] in SiW protocol settings, we use Equal Error Rate (EER) as validation metric for all models to search the threshold to report performance.

Dataset	Proxy Tasks			CASIA				Replay Attack				MSU				All	
Attack Type	D	M	R	V	C-P	W-P	All	V	D-P	P-P	All	P-P	H-V	M-V	All	Mean	Std
OC-SVM$_{\text{RBF}}$+IMQ[50]				63.26	59.43	66.81	63.34	84.48	67.57	70.30	74.49	53.94	84.75	76.56	72.61	70.14	4.87
OC-SVM$_{\text{RBF}}$+BSIF[50]				67.59	51.01	**96.33**	72.76	46.54	63.24	38.88	50.62	62.06	80.56	64.06	69.25	64.21	9.71
SVM$_{\text{RBF}}$+LBP[67]				77.41	**87.14**	69.48	77.61	69.64	73.31	71.85	71.58	55.39	96.02	94.88	83.36	77.51	4.80
NN+LBP[51]				71.80	70.26	67.55	69.78	36.93	75.43	69.45	59.75	26.10	96.84	85.31	71.48	69.75	8.30
GMM+LBP[51]				65.41	85.00	50.15	66.06	60.78	61.46	55.32	59.57	59.35	91.18	86.43	79.92	68.51	8.48
OC-SVM$_{\text{RBF}}$[51]				64.94	85.75	55.15	67.95	84.83	72.62	57.34	73.01	60.90	68.41	75.51	68.60	69.85	2.24
AE+LBP[51]				77.72	80.30	52.92	69.56	79.67	54.92	52.71	63.39	55.67	87.94	92.18	79.67	70.87	6.71
Auxiliary [4]				-	-	-	73.15	-	-	-	71.69	-	-	-	85.88	76.90	6.37
*MADDG [7]				-	-	-	84.51	-	-	-	84.99	-	-	-	88.06	85.85	1.57
Ours				74.88	77.44	81.17	79.31	82.09	72.96	91.42	84.44	66.25	96.60	95.34	85.81	83.18	2.79
	✓			80.02	81.79	87.80	87.61	85.54	84.37	95.65	84.82	67.82	97.52	96.16	88.59	87.00	1.59
	✓	✓		80.43	81.34	89.24	87.80	86.15	84.56	95.63	85.23	68.14	97.54	97.02	88.24	87.09	1.32
	✓	✓	✓	**80.69**	82.13	90.06	**87.92**	**86.69**	**84.92**	**96.33**	**85.27**	**68.23**	**97.70**	**97.50**	**89.23**	**87.47**	1.64

Table V: Inter-dataset evaluation on CASIA, Replay Attack and MSU, AUC(%) is reported. We follow the "Leave one dataset & attack-type out" protocol in [50], where the attack types in testing set is unseen in the training set. We abbreviate V, C-P, W-P, D-P, P-P, H-V and M-V for Video, Cut Photo, Warped Photo, Digital Photo, Printed Photo, HR Video and Mobile video, respectively. Best results are shown in **Bold**. *: retrained by their released codes.

V.C Intra-Dataset Evaluation

We evaluate on a recent large scale spoofing dataset SiW with carefully designed cross attack type testing protocols. We refer another intra-dataset evaluation on Replay-Attack to supplementary due to space limit.

Domain generalization with single source domain In Table VII, We propose a more challenging domain generalization protocol while using only a *single* source domain as the training set. Here we use SiW for training, while CASIA, Replay-Attack, MSU-MFSD and Oulu-NPU for testing. Since it is a new protocol and seldom method conduct this protocol, we implement the baselines of RBF kernel SVM classifier combined with BSIF and LBP feature, respectively.

SiW Evaluation: There are 3 protocols in SiW. Protocol 1 focus on evaluating the performance of variations in face pose and expression. Protocol 2 focuses on the unseen medium of replay attack. It chooses 3 out of 4 display attacks, as training and leaving the remaining one as testing, which is iteratively conducted 4 times and averaged. Protocol 3 evaluates cross presentation attack detection, i.e., from print attack to replay attack and vice versa. Averaging over the two is reported.

In Table II, our method consistently outperforms the other methods with significant margin, i.e., on Protocol 1, we achieve **0.36** ACER while FAS-TD-SF [27] is 1.05. On Protocol 2, ours is **0.20** while the best compared method is 0.31 from FAS-TD-SF. On Protocol 3, ours is **7.32** while the best compared method is 7.73. Similar to Replay-Attack, we apply a gradually increasing module way to highlight effectiveness of the proposed modules. The ablation over our proposed modules suggests: (1) Depth, reflection and material are beneficial proxy tasks. (2) putting more proxy tasks together boosts the performance. (3) Our uncertainty-aware attention on top of the baselines can further achieve performance gain with significant margin. We also show an ablation contrasting w/ or w/o using Attribute-assisted triplet mining, it shows that by adding triplet constraint, there is continuous margin gain over the other baselines.

Replay-Attack Evaluation: In Replay Attack dataset, there are three attack types, namely video, digital photo and printed photo in Replay-Attack. We follow the intra-dataset protocol in [50], utilizing two out of three attack type data for training and the left attack type for testing.

Traditional methods focus on leveraging different classifiers, i.e., one-class SVM and SVM with RBF kernel, combined with different features, such as LBP and BSIF. Ablation study show our methods outperforms both traditional methods and recent deep learning based method DTL [5] even with incomplete components.

V.D Inter-Dataset Evaluation

The inter-dataset setting mimics the real setting for unseen attack across types and datasets. We consider several protocols. (a) Directly transfer from the attributes-rich CelebA-Spoof[11] dataset to CASIA dataset (b) The traditional "Leave one dataset & attack-type out" protocol[50], taking CASIA, Replay-Attack and MSU-MFSD as our datasets. Each of the three datasets contains three attack types. When evaluating one attack type of one dataset, we pick the

Train → Test Dataset	DD	Proxy Tasks			OCI→M		OMI→C		OCM→I		ICM→O	
Method		D	M	R	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MS LBP[72]	×				29.76	78.50	54.28	44.98	50.30	51.64	50.29	49.31
Binary CNN[73]					29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
IDA [70]					66.67	27.86	55.17	39.05	28.35	78.25	54.20	44.59
Color Texture [74]					28.09	78.47	30.58	76.89	40.40	62.78	63.59	32.71
LBPTOP [75]					36.90	70.80	42.60	61.05	49.45	49.54	53.15	44.09
Auxiliary [4]					22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
MADDG [76]	✓				17.61	88.06	24.50	84.51	22.19	84.99	27.98	80.02
RFGML [77]					**13.89**	**93.98**	20.27	88.16	17.30	90.48	**16.45**	**91.16**
SSDG-M [78]					16.67	90.47	23.11	85.45	18.21	**94.61**	25.17	81.83
Ours	×				23.25	85.81	28.13	79.31	22.25	84.44	32.51	76.26
		✓			17.26	88.59	20.21	87.61	22.46	86.74	28.94	80.30
		✓	✓		16.68	90.73	18.57	89.03	19.78	88.42	23.76	83.58
		✓	✓	✓	15.68	91.32	**18.39**	**89.28**	**17.21**	91.83	21.93	85.48

Table VI: Inter-dataset evaluation on CASIA, Replay Attack, MSU and Oulu-NPU dataset. AUC (%) and HTER (%) is reported. We follow the "Leave one dataset out" protocol in [7], where training and testing sets share attack types. **O,C,I,M** denote Oulu-NPU[67], CASIA[36], Replay-Attack[69], MSU-MSFD[70] dataset, respectively. **DD** denotes disentangling source domains. Best results are shown in **Bold**.

Train → Test Dataset	Proxy Tasks			SiW → CASIA			SiW → Replay Attack			SiW → MSU			SiW → Oulu-NPU		Average	
Attack Type	D	M	R	V	C-P	W-P	Video	D-P	P-P	P-P	H-V	M-V	V	P-P	Mean	Std
SVM_{RBF}+BSIF [51]				77.12	79.38	78.27	79.93	43.82	76.53	50.01	90.13	89.54	69.77	80.84	74.12	13.98
SVM_{RBF}+LBP[67]				80.00	83.08	80.74	82.24	54.94	80.89	66.59	98.76	**97.15**	73.31	85.94	80.33	11.86
Ours				87.64	82.40	81.56	91.68	81.24	72.87	73.86	91.02	89.61	81.08	81.72	83.15	6.04
	✓			88.23	83.47	81.68	91.83	82.23	79.83	75.34	93.83	89.73	81.87	82.49	84.59	5.33
	✓	✓		90.03	85.64	82.03	92.21	82.47	81.24	81.93	97.78	92.31	82.53	86.94	86.82	5.27
	✓	✓	✓	**90.25**	**86.17**	**82.31**	**93.06**	**82.97**	**83.48**	**85.86**	**98.82**	93.27	**89.62**	**87.52**	**88.48**	4.87

Table VII: Domain generalization with single source domain, we show result from SiW to CASIA, Replay Attack and MSU, Oulu-NPU, respectively. AUC(%) is reported. Best results are shown in **Bold**.

other two datasets for training and excluding the testing attack type from training. (c) A less-strict "Leave one dataset out" protocol used in MADDG[7], the difference to (b) is that, in this protocol training and testing sets would have overlapping attack types. (d) A more challenging domain generalization protocol while only a ***single*** source domain is allow to use as the training set.

CelebA-Spoof Evaluation In Table IV, we conduct a ablation on our Attribute-assisted Triplet Mining utilizing face attributes introduced in CelebA. We compare with the current SoTA AENet[11] which utilize extra task such as Illumination Conditions and Spoof Type Classification. However, without the extra information, our methods still consistently outperform AENet. By adding our Attribute-assisted Triplet Mining, we further boost the performance further by 1.2% in terms of HTER.

Leave one dataset & attack-type out Evaluation In Table V, we evaluate each of the three attack types from three datasets. Both traditional feature learning based methods and most recent deep learning based methods [4, 7] are compared. Overall we achieve consistently stronger results than the other methods. In CASIA, video attack is significantly better than other methods while cut photo and warped photo are among the top. In Replay-Attack, we achieve clear better performance. In MSU-MFSD, we observe 1% to 6% performance improvement over the compared methods.

Leave one dataset out Evaluation In Table VI. Since our physically-guided proxy task does not require any domain priors, we compare methods w and w/o source domains disentanglement (DD). Our method surpass all methods without domain disentanglement including [4], while still achieve comparable performance compare to methods [76][77][78] that utilize extra source domains information.

V.E Analysis of Uncertainty-aware Attention

We visualize our Uncertainty-aware Attention Modeling in Figure 5 and 6. Specifically, we visualize the last activation map before average pooling across Liveness, Depth, Reflection and Material channel alongside with the corresponding Input Image, Ground-Truth/Predicted Depth map and Ground-Truth/Predicted Reflection map. We also show their respective average value $\mu_{\mathbf{u}}$ on the top of each attention map and depth/reflection map. For spoof face images in Figure 5, we observe that the material attention map focus more on human skin, indicating its strong correlation with material, while the liveness attention map focus on screen and background texture. For genuine face images Figure 6, the liveness attention map lean heavily on face itself, while the reflection and material attention map have more weight on the edge of faces and the depth attention map focus on hair and eyes.

VI Conclusion

In this work, we propose depth, reflection and material guided proxy tasks for unseen spoofing attacks. We propose a multi-source multi-channel training scheme for model optimization. Due to the consistency of depth, reflection and skin material across different spoofing scenario on genuine faces, by harnessing those physical proxy tasks, we expect the proposed method to deal with unseen spoofing attacks. Finally, an uncertainty-aware attention module is introduced to aggregate the multiple channels for boosted performance. Experiments across intra- and inter-dataset protocols show our method achieves consistently better performance and is effective for unseen spoofing detection.

Declarations

- Funding: Not Applicable
- Conflict of interest/Competing interests: Not Applicable
- Ethics approval: Not Applicable
- Consent to participate: Not Applicable
- Consent for publication: Not Applicable
- Availability of data and materials: All mentioned datasets, including CASIA [68], Replay-Attack [69], MSU-MFSD [70], Oulu-NPU [67], SiW [4], CelebA-Spoof [11] are publicly available datasets.
- Code availability: Custom proprietary code.
- Authors' contributions: Junru Wu contributed to idea development, code implementation and writing of the paper, Xiang Yu contributed to idea development, code implementation and writing of the paper, Buyu Liu contributed to code implementation and writing of the paper, Zhangyang Wang and Manmohan Chandraker contributed to writing of the paper.

References

[1] Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH (1999)

[2] Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material recognition in the wild with the materials in context database. In: CVPR (2015)

[3] Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4786–4794 (2018)

[4] Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: CVPR (2018)

[5] Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X.: Deep tree learning for zero-shot face antispoofing. In: CVPR (2019)

[6] Patel, K., Han, H., Jain, A.K.: Cross-database face anti-spoofing with robust feature representation. In: CCBR (2016)

[7] Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: CVPR (2019)

[8] Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., Zhou, F., Lei, Z.: 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 (2020)

[9] Zhang, K.-Y., Yao, T., Zhang, J., Liu, S., Yin, B., Ding, S., Li, J.: Structure destruction and content combination for face anti-spoofing. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–6 (2021). IEEE

[10] Yu, Z., Li, X., Shi, J., Xia, Z., Zhao, G.: Revisiting pixel-wise supervision for face anti-spoofing.

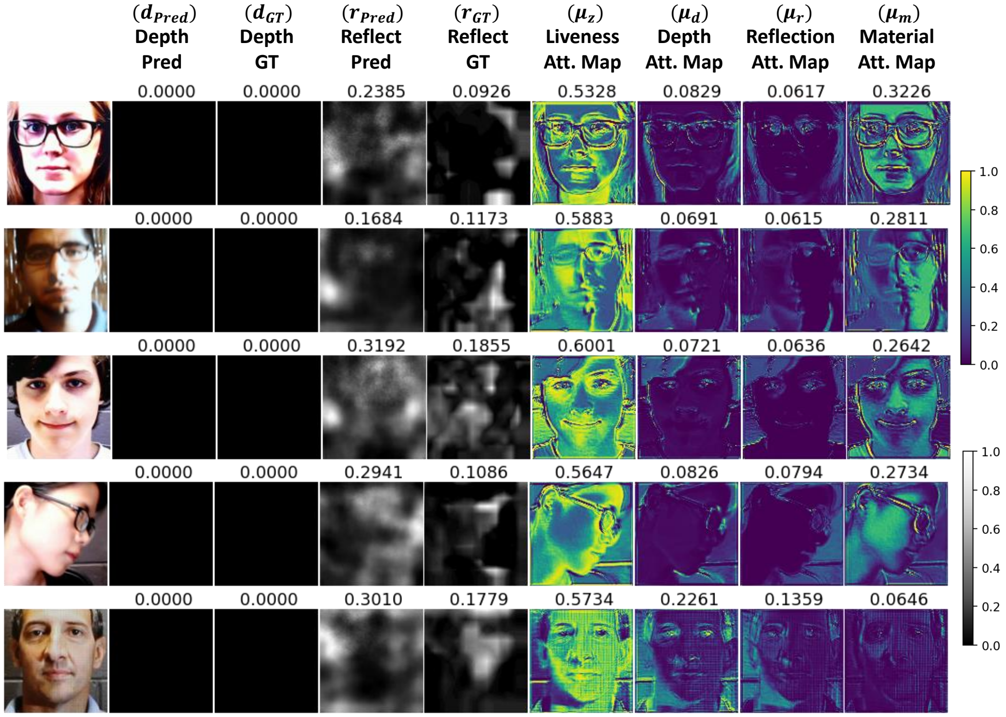


Fig. 5: Visualization of our Uncertainty-aware Attention Modeling on **spoof** faces in SiW dataset. We show the input image alongside with GT/predicted depth maps, GT/predicted reflection maps. We also show our uncertainty-aware attention maps of Liveness, Depth, Reflection and Material channel, where we visualize the last activation map bef

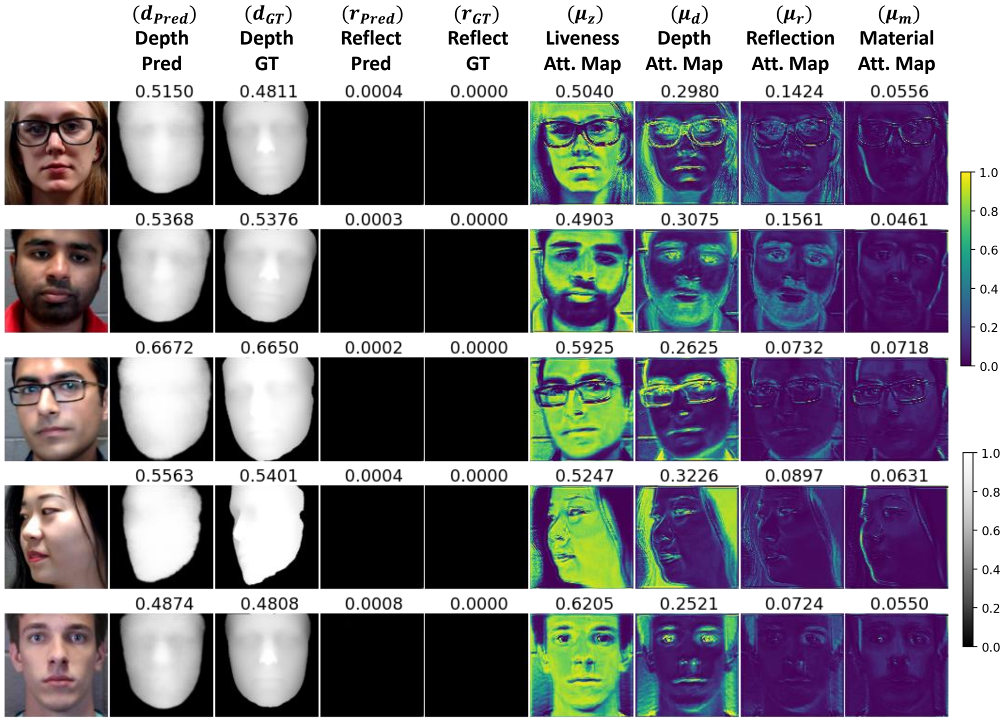


Fig. 6: Visualization of our Uncertainty-aware Attention Modeling on **genuine** faces in SiW dataset. We show the input image alongside with GT/predicted depth maps, GT/predicted reflection maps. We also show our uncertainty-aware attention maps of Liveness, Depth, Reflection and Material channel, where we visualize the last activation map before average pooling.

IEEE Transactions on Biometrics, Behavior, and Identity Science (2021)

[11] Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., Liu, Z.: Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In: European Conference on Computer Vision, pp. 70–85 (2020). Springer

[12] Yu, Z., Li, X., Niu, X., Shi, J., Zhao, G.: Face anti-spoofing with human material perception. In: European Conference on Computer Vision, pp. 557–575 (2020). Springer

[13] Kollreider, K., Fronthaler, H., Faraj, M.I., Bigun, J.: Real-time face detection and motion analysis with application in liveness assessment. TIFS (2007)

[14] Pan, G., Sun, L., Wu, Z., Lao, S.: Eyeblink based anti-spoofing in face recognition from a generic webcamera. In: ICCV (2007)

[15] de Haan, G., Jeanne, V.: Robust pulse rate from chrominance-based rppg. IEEE Trans. Biomedical Engi- neering (2013)

[16] Bobbia, S., Benezeth, Y., Dubois, J.: Remote photo-plethysmography based on implicit living skin tissue segmentation. In: ICPR (2016)

[17] Liu, S., Yuen, P.C., Zhao, S.Z.G.: 3d mask face anti-spoofing with remote photoplethysmography. In: ECCV (2016)

[18] Nowara, E.M., Sabharwal, A., Veeraraghavan, A.: Ppgsecure: Biometric presentation attack detection using photo-pletysmograms. In: FG (2017)

[19] Kim, T., Kim, Y., Kim, I., Kim, D.: Basn: Enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 0–0 (2019)

[20] Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: ECCV (2018)

[21] Komulainen, J., Hadid, A., Pietikainen, M.: Context based face anti-spoofing. In: BATS (2013)

[22] J.Yang, Z.Lei, S.Liao, S.Z.Li: Face liveness detection with component dependent descriptor. In: ICB (2013)

[23] Patel, K., Han, H., Jain, A.K.: Secure face unlock: Spoof detection on smartphones. TIFS (2016)

[24] Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: ICIP (2015)

[25] Pereira, T., Anjos, A., DeMartino, J.M., Marcel, S.: Lbp-top based counter measure against face spoofing attacks. In: ACCV (2012)

[26] Määttä, J., Hadid, A., Pietikäineninen, M.: Face spoofing detection from single images using micro-texture analysis. In: IJCB (2011)

[27] Wang, Z., Zhao, C., Qin, Y., Zhou, Q., Lei, Z.: Exploiting temporal and depth information for multi-frame face anti-spoofing. arXiv preprint arXiv:1811.05118 (2018)

[28] Zhang, S., Wang, X., Liu, A., Zhao, C., Wan, J., Escalera, S., Shi, H., Wang, Z., Li, S.Z.: A dataset and benchmark for large-scale multi-modal face anti-spoofing. In: CVPR (2019)

[29] Li, J., Wang, Y., Tan, T., Jain, A.K.: Live face detection based on the analysis of fourier spectra. In: SPIE (2004)

[30] Tan, X., Li, Y., Liu, J., Jiang, L.: Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: ECCV (2010)

[31] Agarwal, A., Singh, R., , Vatsa, M.: Face anti-spoofing using haralick features. In: BATS (2016)

[32] Bao, W., Li, H., Li, N., Jiang, W.: A liveness detection method for face recognition based on optical flow field. In: IASP (2009)

[33] Siddiqui, T.A., Bharadwaj, S., Dhamecha, T.I., Agarwal, A., Vatsa, M., Singh, R., Ratha, N.: Face anti-spoofing with multifeature videolet

aggregation. In: ICPR (2016)

[34] Feng, L., Po, L., Li, Y., Xu, X., Yuan, F., Cheung, T.C., Cheung., K.: Integration of image quality and motion cues for face anti-spoofing: A neural network approach. Journal of Visual Communication and Image Representation (2016)

[35] Z.Xu, S.Li, W.Deng: Learning temporal features using lstm-cnn architecture for face anti-spoofing. In: ACPR (2016)

[36] Yang, Z., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. In: arXiv:1408.5601 (2014)

[37] Li, L., Feng, X., Boulkenafet, Z., Xia, Z., Li, M., Hadid, A.: An original face anti-spoofing approach using partial convolutional neural network. In: IPTA (2016)

[38] Atoum, Y., Liu, Y., Jourabloo, A., Liu, X.: Face antispoofing using patch and depth-based cnns. In: IJCB (2017)

[39] Jourabloo, A., Liu, Y., Liu, X.: Face de-spoofing: Anti-spoofing via noise modeling. In: ECCV (2018)

[40] Cai, R., Li, H., Wang, S., Chen, C., Kot, A.C.: Drl-fas: a novel framework based on deep reinforcement learning for face anti-spoofing. IEEE Transactions on Information Forensics and Security **16**, 937–951 (2020)

[41] Chen, B., Yang, W., Li, H., Wang, S., Kwong, S.: Camera invariant feature learning for generalized face anti-spoofing. IEEE Transactions on Information Forensics and Security **16**, 2477–2492 (2021)

[42] Liu, A., Tan, Z., Wan, J., Liang, Y., Lei, Z., Guo, G., Li, S.Z.: Face anti-spoofing via adversarial cross-modality translation. IEEE Transactions on Information Forensics and Security **16**, 2759–2772 (2021)

[43] Di Martino, J.M., Qiu, Q., Sapiro, G.: Rethinking shape from shading for spoofing detection. IEEE Transactions on Image Processing **30**, 1086–1099 (2020)

[44] Yu, Z., Qin, Y., Zhao, H., Li, X., Zhao, G.: Dual-cross central difference network for face anti-spoofing. arXiv preprint arXiv:2105.01290 (2021)

[45] Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5295–5305 (2020)

[46] Yu, Z., Wan, J., Qin, Y., Li, X., Li, S.Z., Zhao, G.: Nas-fas: Static-dynamic central difference network search for face anti-spoofing. arXiv preprint arXiv:2011.02062 (2020)

[47] Quan, R., Wu, Y., Yu, X., Yang, Y.: Progressive transfer learning for face anti-spoofing. IEEE Transactions on Image Processing **30**, 3946–3955 (2021)

[48] Liu, S., Zhang, K.-Y., Yao, T., Bi, M., Ding, S., Li, J., Huang, F., Ma, L.: Adaptive normalized representation learning for generalizable face anti-spoofing. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1469–1477 (2021)

[49] Liu, S., Zhang, K.-Y., Yao, T., Sheng, K., Ding, S., Tai, Y., Li, J., Xie, Y., Ma, L.: Dual reweighting domain generalization for face presentation attack detection. arXiv preprint arXiv:2106.16128 (2021)

[50] Arashloo, S.R., Kittler, J., Christmas, W.: Anomaly detection approach to face spoofing detection: a new formulation and evaluation protocol. IEEE Access (2017)

[51] Xiong, F., AbdAlmageed, W.: Unknown presentation attack detection with face rgb images. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–9 (2018). IEEE

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

[53] Qin, Y., Zhao, C., Zhu, X., Wang, Z., Yu, Z., Fu, T., Zhou, F., Shi, J., Lei, Z.: Learning meta model for zero-and few-shot face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11916–11923 (2020)

[54] Qin, Y., Yu, Z., Yan, L., Wang, Z., Zhao, C., Lei, Z.: Meta-teacher for face anti-spoofing. IEEE transactions on pattern analysis and machine intelligence (2021)

[55] Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Machine Learning Proceedings 1994, pp. 148–156. Elsevier, ??? (1994)

[56] Sun, Q., Laddha, A., Batra, D.: Active learning for structured probabilistic models with histogram approximation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3612–3621 (2015)

[57] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems, pp. 5574–5584 (2017)

[58] Mukhoti, J., Gal, Y.: Evaluating bayesian deep learning methods for semantic segmentation. arXiv preprint arXiv:1811.12709 (2018)

[59] Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2372–2379 (2009). IEEE

[60] Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7482–7491 (2018)

[61] Betta, G., Capriglione, D., Liguori, C., Paolillo, A.: Uncertainty evaluation in face recognition algorithms. In: 2011 IEEE International Instrumentation and Measurement Technology Conference, pp. 1–6 (2011). IEEE

[62] Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: CVPR (2014)

[63] Lee, C.Y., Xie, S., Gallagher, P.W., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS (2015)

[64] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)

[65] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)

[66] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)

[67] Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: Oulu-npu: A mobile face presentation attack database with real-world variations. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 612–618 (2017). IEEE

[68] Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: 2012 5th IAPR International Conference on Biometrics (ICB), pp. 26–31 (2012). IEEE

[69] Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face antispoofing. In: 2012 BIOSIG-proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG), pp. 1–7 (2012). IEEE

[70] Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security **10**(4), 746–761 (2015)

[71] ISO Central Secretary: Information technology — biometric presentation attack detection — part 1: Framework. Standard ISO/IEC 30107-1:2016, International Organization for Standardization, Geneva, CH (2016). https://www.iso.org/standard/53227.html

[72] Määttä, J., Hadid, A., Pietikäinen, M.: Face spoofing detection from single images using micro-texture analysis. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–7 (2011). IEEE

[73] Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601 (2014)

[74] Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. IEEE Transactions on Information Forensics and Security **11**(8), 1818–1830 (2016)

[75] de Freitas Pereira, T., Komulainen, J., Anjos, A., De Martino, J.M., Hadid, A., Pietikäinen, M., Marcel, S.: Face liveness detection using dynamic texture. EURASIP Journal on Image and Video Processing **2014**(1), 2 (2014)

[76] Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10023–10031 (2019)

[77] Shao, R., Lan, X., Yuen, P.C.: Regularized fine-grained meta face anti-spoofing. In: AAAI, pp. 11974–11981 (2020)

[78] Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8484–8493 (2020)