BASN: Enriching Feature Representation Using Bipartite Auxiliary Supervisions for Face Anti-Spoofing

Taewook Kim[1], Yonghyun Kim[1,2], Inhan Kim[1], Daijin Kim[1]
[1]POSTECH [2]Kakao Corp.
{taewook101, kiminhan, dkim}@postech.ac.kr, aiden.kyh@kakaocorp.com

Abstract

Face anti-spoofing is an important task to assure the security of face recognition systems. To be applicable to unconstrained real-world environments, generalization capabilities of the face anti-spoofing methods are required. In this work, we present a face anti-spoofing method with robust generalization ability to unseen environments. To achieve our goal, we suggest bipartite auxiliary supervision to properly guide networks to learn generalizable features. We propose a bipartite auxiliary supervision network (BASN) that comprehensively utilizes the suggested supervision to accurately detect presentation attacks. We evaluate our method by conducting experiments on public benchmark datasets and we achieve state-of-the-art performances.

1. Introduction

Current state-of-the-art face recognition methods [11, 12, 22, 34] can recognize faces almost perfectly with accuracy that exceeds what humans can achieve. However, these face-recognition technologies may give false authentication when the system is presented with spoof images, such as video replays or photographs. Since face spoofs are easier to acquire compared to other biometric modalities (e.g., fingerprints or iris), face recognition systems can be easily fooled, and therefore, fragility exists in the face recognition system. To assure the security of face recognition, face anti-spoofing methods to prevent these deceptions must be developed.

Several methods to prevent face spoofing have been developed. Some methods used Local Binary Patterns (LBPs) [8, 23] to capture textural differences between presentation attacks (PAs) and live faces. Another method [14] attempted to utilize moiré-pattern to distinguish between real faces and printed photographs with analysis in Fourier spectra. Some other methods [25, 38] used motion cues such as lip movement and eye blinking to distinguish real faces from PAs. However, these methods that rely on handcrafted features do not achieve powerful feature representation and show limited performance that decreases in cross-tests [3, 5, 10, 41].

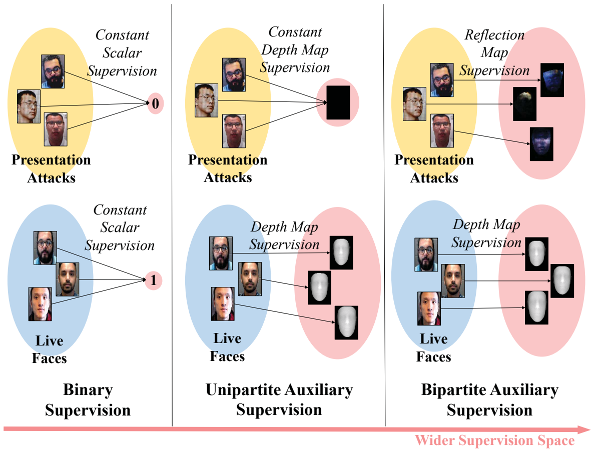

Figure 1. Comparison of face anti-spoofing using binary, unipartite and bipartite auxiliary supervision. Contours of input spaces of spoof (yellow) and live (blue) images, and supervision space (red). Compared to the simple binary supervision and unipartite supervision, bipartite auxiliary supervision helps to learn mapping relation from input space to a wider supervision space. In this way, feature representation can be enriched, and models can be guided to capture more generalizable features of both live faces and presentation attacks.

Recently, deep neural networks have achieved great success in computer vision society by outperforming former state-of-the-art performances in almost every task [15, 18, 33, 45], including face anti-spoofing [21, 28, 43, 46, 49]. [28] fine-tuned a CNN pre-trained on ImageNet [42], then used an SVM-based classifier to detect PAs. [46] adopted LSTM-CNN architecture and attempted to learn both spatial information and temporal patterns of the input sequence.

Despite the progress in performance, existing methods

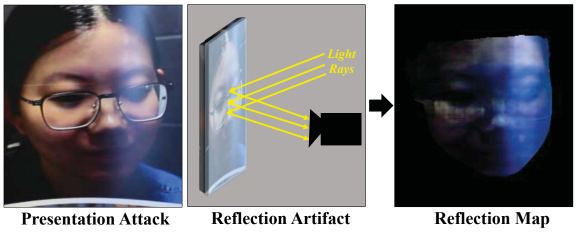

Figure 2. Reflection map as auxiliary supervision. Light rays that travel from the photographers side to the surfaces of PAs (e.g., coatings of photographs) are reflected and captured by cameras, and thereby cause reflection artifacts. Bipartite auxiliary supervision is achieved by using this reflection artifact along with facial depth maps as auxiliary supervision.

are not readily generalized, and show great degradation of performance when evaluated on datasets that they are not trained on. Most of the deep learning-based methods guide networks with binary supervision that only uses binary class labels with softmax loss [28, 39, 47]. These methods that only use class labels as supervision do not properly guide networks to capture generalizable features, and the networks are prone to capture arbitrary features that may only exist in the train set. As a result, networks are easily get overfitted and lead to deterioration of performances.

To solve this problem, [32] provided auxiliary 3D facial depth map and remote-photoplethysmography (r-ppg) signals to serve as supervision for the network. They exploited the fact that spoof media do not exhibit face-like depth, nor physiological signals. The provided information of facial depth map and r-ppg signal served as additional ground truth for the network, and the network was trained to predict depth maps and physiological signals. As a result, the network was trained to better capture generalizable features and made notable improvements in performance.

However, the auxiliary supervision that was used in [32] only has faithful meaning for the live face and is less meaningful for PAs because the provided supervision is constant-valued (i.e., constant depth map or constant r-ppg signal for PAs). Similar to the simple binary supervision, this type of supervision can mislead the neural network to learn arbitrary features of PAs that may not be generalizable on a test set. As a result, error rates increase, and generalization capability is degraded. Also, this can limit the feature representation of a network because the supervision guides the network to learn a mapping relation from input space to a limited supervision space (Fig. 1). Therefore, the supervision of spoof images requires increased meaning and complexity that help to capture generalizable features.

We propose a bipartite auxiliary supervision network (BASN) that comprehensively utilizes the following faithful auxiliary information of both spoofs and live images: 3D facial depth map for live images and spoof medium reflection map for spoof images. Light rays that are reflected from a surface of spoof medium cause reflection artifacts in recaptured images (Fig. 2). Compared to spoof media, surfaces of live faces show differences in smoothness and textures, and reflection artifacts rarely occur during the image capturing. Therefore, if we can accurately capture reflection artifacts, spoofs can be effectively detected. We thereby aim to estimate the reflection map by providing auxiliary reflection supervision. Along with the reflection supervision, we also use faithful depth map information as auxiliary supervision for live faces. In this way, meaningful supervision can be utilized for both live and spoof images. We call this type of supervision as bipartite auxiliary supervision since we use meaningful supervision for both spoof and live images. We type the former method [32] that utilizes meaningful supervision for either spoof or live images as unipartite auxiliary supervision. The proposed BASN achieves robust generalization ability by extracting discriminative and generalizable features of both spoof and live images and by comprehensively interpreting them.

Contributions of our work are summarized as follows:

- We exploit a depth map and a reflection map as bipartite auxiliary supervision that well represents characteristics of both live and spoof faces. Bipartite auxiliary supervision helps to enrich feature representation and improve the generalization capabilities of models.
- We propose BASN, an architecture that effectively learns to extract and fuse auxiliary features to detect presentation attacks.
- We evaluate our model on several publicly-available face anti-spoofing datasets and achieve comparable to or higher performance than the state-of-the-art methods.

2. Related Works

Initially, researchers mainly focused on hand-crafted feature representation, such as LBP [5, 9, 10, 36], HoG [26, 48] and SIFT [40] to tackle the problem. However, these hand-crafted feature-based methods showed a lack of generalization capability due to limited feature representation. After the great success of deep learning in several computer vision tasks [15, 18, 33, 45], researchers have more focused on approaches based on CNNs. [47] first attempted to utilize deep features for face anti-spoofing with Alexnet [27] architecture. [39] proposed to learn textural features from both facial and non-facial regions under CNN framework, believing that spoof patterns exist in the whole frame. They also attempted to detect eye blinking by estimating frame difference and utilized the state change of eye movements as an additional clue. [28, 37] approached to apply SVM as a classifier with deep features extracted by CNN. [28] proposed to extract partial convolutional features by gathering thresholded values of feature maps and applied PCA to extract generalizable features. In [29], LBP features are extracted with CNNs, and the extracted features are utilized to distinguish between spoofs and live images. Also, there were some attempts to capture both spatial and temporal relation of the input sequence by adopting LSTM-CNN structure [46, 49]. [49] approached to learn spatio-temporal features by attending to discriminative regions of input images, and made contributions with an additional data synthesis method.

Recently, there were several attempts of utilizing auxiliary supervision to guide networks, aiming to avoid overfitting. [2] proposed to extract both local features of patches and depth maps of facial regions by adopting two-stream CNN architecture. [43] attempted to achieve improvements in generalization capabilities by adopting an adversarial-learning [16] strategy that a generator aims to fool multiple adversarial discriminators pre-trained on different domains. [21] regarded face anti-spoofing as image decomposition problem and attempted to estimate spoof noise pattern with auxiliary supervision. [32] proposed to guide networks with auxiliary supervision of facial depth information and remote-photoplethysmography (r-ppg) signal. The methods proposed to utilize auxiliary supervision were more properly guided to capture generalizable features and achieved notable improvements in PA detection performance.

3. The Proposed Method

The main goal of this work is to guide the network with faithful information of both spoof and live images; We aim to utilize 3D facial depth map supervision for live faces and spoof medium reflection map supervision for spoof images. Similar to simple binary supervision, training networks with unipartite auxiliary supervision (i.e., supervision with faithful information for either live or spoof images) lead to degradation of generalization capability. Therefore, instead of providing meaningful supervision only for live faces, we also provide reflection maps that well represent the characteristics of face spoofs to serve as auxiliary supervision. To realize our goal, we propose a bipartite auxiliary supervision network (BASN) (Fig. 3) that comprehensively utilizes the suggested bipartite auxiliary supervision. The proposed BASN extracts faithful features of both spoof and live images with bipartite auxiliary supervision, and enriches the extracted features for effective counter-measurement against face spoofing.

3.1. Bipartite Auxiliary Supervision

Depth Map. We use 3D facial depth map information to detect PAs. Spoof medium and live face show distinct differences in shapes. Spoof media exhibit even and flat surfaces, whereas live faces show more irregular shapes that are not flat. This key difference in appearance can be represented as a depth map, and we thereby aim to train the network to estimate depth maps to detect PAs. Since there is no ground truth label of the depth map in spoof datasets, we estimate depth maps of face images by using the existing dense-face-alignment method [13] and regard them as the ground truths. Instead of estimating the actual depth of the image, the ground truth depth map is provided by estimating 3D shapes of a face, and the estimated depth values are limited to the facial region. The ground truth facial depth map D is defined as,

$$D(I|y) = \begin{cases} 0, & \text{if y is spoof,} \\ \frac{1}{|\mathrm{D}|}\widetilde{d}(I), & \text{if y is live,} \end{cases} \tag{1}$$

where I is a given image, y is a label of I and $\widetilde{d}$ is a depth map of a face. The values of depth map are normalized to the range $[0, 1]$ by using the normalization factor $|\mathrm{D}|$, where 0 indicates farthest from the viewer and 1 indicates closest to the viewer. Following the suggestion from [32], we use facial depth map consisting only of zeros as supervision for fake images, assuming that spoof media are on flat structures.

Reflection Map. Using meaningful supervision only for live images (i.e., unipartite auxiliary supervision) is not enough for the network to capture faithful features of PAs, because the constant map supervision does not consider spatial patterns. We use a reflection map as additional auxiliary supervision to remedy this lack. As the reflection artifacts caused by reflected lights from smooth and flat surfaces of spoof media rarely occur from live faces, these artifacts are key difference that help to distinguish spoof from live faces. We thereby aim to capture reflection artifacts of spoof images by utilizing reflection maps that represent the artifacts as auxiliary supervision. By allowing reflection map to serve as additional auxiliary supervision, bipartite auxiliary supervision is achieved (i.e., faithful supervision for both live and spoof images), and the network is guided to learn more generalizable features of both PAs and live images. As a result, accurate PA detection can be accomplished. We define the ground truth reflection map R as auxiliary supervision as,

$$R(I|y) = \begin{cases} \frac{1}{|\mathrm{R}|}\widetilde{r}(I), & \text{if y is spoof,} \\ 0, & \text{if y is live,} \end{cases} \tag{2}$$

where I is a given image, y is a label of I and $\widetilde{r}$ is an estimated reflection artifact by [51]. Since the ground truth data of reflection map $\widetilde{r}$ is not provided in spoofing datasets, we estimate the reflection map by using the state-of-the-art reflection estimation network [51] and regard them as the ground truths. The generated reflection maps have 3 channels for RGB whereas depth map only has a single channel.

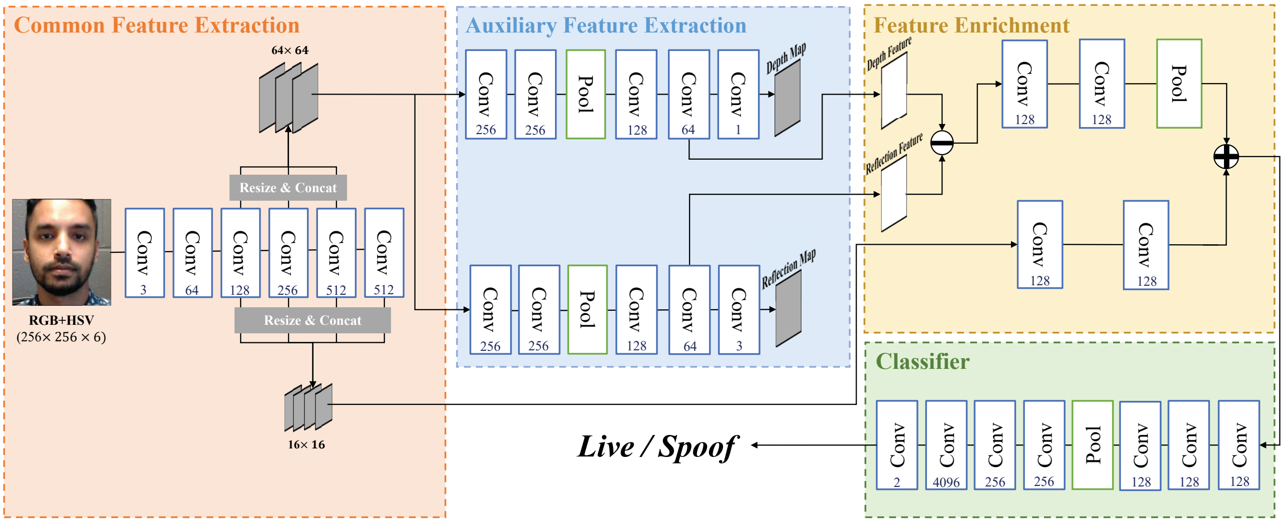

Figure 3. The proposed architecture of BASN. The number of kernels is denoted in the bottom of each convolutional layer. The kernel size of the last two convolutional layers of the classifier is 1×1 and 8×8, respectively. Every other convolutional layer of BASN has a kernel size of 3×3. Every convolutional layer uses a stride of 1. The filter size is 2×2 for pooling layers with strides of 2.

For consistency with depth map supervision, and to avoid being overfitted to backgrounds, we crop face regions from reflection maps. To crop face regions, facial depth maps are estimated by using the dense-face-alignment-network [13], and the depth maps are thresholded to generate a binary mask. Then, the inner product of the binary mask and the reflection map is computed, and only the face region of reflection map remains. Also, each RGB value of the reflection maps is divided by the maximum values |R| of the region of each channel, so that the values are normalized to be in the range of $[0, 1]$. As reflection artifacts rarely occur from live faces, we use reflection maps that consist only of zeros for live faces, assuming that live face images do not have reflection artifacts.

3.2. Network Architecture

Backbone Network. The proposed BASN is based on VGG-16 [44] network that is pre-trained on ImageNet [42] (Fig. 3). We use an additional convolutional layer before the first VGG-16 layer that outputs a 3-channel feature map. We use the pre-trained model to extract features of spoofing datasets that are more generalizable. The extracted feature maps are common convolutional features that are shared by depth and reflection feature extractors of auxiliary feature extractor (AFE) module. Motivated by [31], we adopt Feature Pyramid Network (FPN) structure to construct feature maps of multi-scales. Feature maps of conv2, conv3, conv4 from VGG-16 are resized to the fixed size of 64×64, and the feature maps are concatenated to be passed to the AFE. Also, feature maps of conv2, conv3, conv4, and conv5 of VGG-16 are resized to the size of 16×16 and are concatenated as a feature map. The second concatenated feature maps are passed directly to the feature enrichment part (Fig. 3).

Auxiliary Feature Extractor. Auxiliary feature extractor (AFE) receives feature maps from the backbone network to estimate auxiliary maps. Each depth and reflection feature extractors of AFE are identical in structure except for the number of kernels in the last layer of each extractors that outputs auxiliary maps of different number of channels (Fig. 3). Every convolutional layer is followed by a batch normalization layer [19] and ReLU. Each auxiliary feature extractor is supervised to minimize following loss function with the ground truth auxiliary maps:

$$\mathcal{L}_{depth} = \frac{1}{N}\sum_{i=1}^{N} ||M_{depth}(f_i) - D(I_i|y_i)||_1^2 \qquad (3)$$

$$\mathcal{L}_{ref} = \frac{1}{N}\sum_{i=1}^{N} ||M_{ref}(f_i) - R(I_i|y_i)||_1^2 \qquad (4)$$

where N is the batch size, M is predicted auxiliary maps of AFE module and f_i denotes concatenated feature maps from the backbone network. The second to last feature maps of each depth and reflection feature extractors are passed to following layers of feature enrichment.

Feature Enrichment and Classification. Feature representations are enriched by fusing feature maps from AFE and the backbone network. A reflection feature map is subtracted feature-wise from a depth feature map (Fig. 3). After subtraction, two consecutive convolutional layers and a max-pooling layer follow. The feature map from the backbone network is fed to convolutional layers and the output

Category	Method	APCER (%)	BPCER (%)	ACER (%)
Baseline	Baseline	7.1	12.5	9.8
Supervision	Model A	**1.0**	14.2	7.6
	Model B	4.3	14.2	9.3
Enrichment	Model C	6.7	10.0	8.3
	Model D	1.5	5.8	**3.6**

Table 1. The ablation study results on Oulu-NPU Protocol 1. Each Model A and B is trained with unipartite supervision of depth map and reflection map, respectively. Model C and Model D are trained with bipartite supervision with different feature fusion method. The depth and reflection feature map of Model C is feature-wise added, and Model D is the proposed BASN.

feature map is feature-wise added to the subtracted feature map. To maintain negative values of feature maps, every convolutional layer is not followed by activation functions for feature fusion. Also, to prevent certain feature maps from dominating the fusing process, a batch normalization layer that helps to balance the scale of a feature map follows each of the convolutional layers. Finally, the fused feature map with enriched feature representation is passed to the classifier to make a final decision, and the classifier is supervised to minimize softmax cross-entropy loss.

Loss Functions. The overall training process of our network is conducted in a multi-task learning fashion:

$$\mathcal{L}_{BASN} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{depth} + \lambda_3 \mathcal{L}_{ref} \tag{5}$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{depth}$, and $\mathcal{L}_{ref}$ denote softmax cross-entropy loss function for the classifier, and loss functions for auxiliary map extractors in AFE, respectively. Each λ denotes a weight of each loss functions. We set values for $\lambda_1 = 10$, and $\lambda_2 = \lambda_3 = 0.00005$. The weight values are empirically selected to balance the contribution of each loss term.

4. Experiments

4.1. Datasets and Metrics

We evaluate the performance of our method by evaluating it on four publicly-available datasets: SiW [32], Oulu-NPU [7], CASIA-FASD [52] and Replay-Attack [8]. We conduct intra-testing on Oulu and SiW datasets, and cross-testing by training on a train set of Replay-Attack and testing on a test set of CASIA-FASD datasets, and vice-versa.

Oulu-NPU. This dataset contains 4,950 videos, 990 of which are real, and 3960 of which are spoofs. Four protocols are defined in the dataset, and each protocol evaluates different generalization capabilities. Protocol 1 evaluates generalization capability on backgrounds and illumination. Protocol 2 evaluates generalization capability by training and testing on different types of print- and video-attack. Protocol 3 evaluates generalization capability on types of image capturing devices. Protocol 4 evaluates generalization capability of combinations of all variations that are included in protocols 1-3.

SiW. This dataset contains 4,478 high-resolution videos of 165 individuals. The videos are recorded with a range of factors such as illumination and attack medium. Three different protocols are defined in the dataset and each protocol evaluates a different generalization capability. Protocol 1 evaluates generalization to pose and expression variations. Protocol 2 evaluates generalization to cross-medium of replay-attack. Protocol 3 evaluates generalization to unknown presentation attacks by training and testing on a dataset that exclusively contains videos of replay-attack or print-attack.

CASIA-FASD and Replay-Attack. CASIA-FASD dataset contains 150 real and 450 spoof videos that record 50 individuals under varied illumination and resolution. Replay-Attack includes 1200 videos of 50 individuals that are provided in a single resolution. CASIA-FASD and Replay-Attack datasets are used for cross-testing.

Metrics. We report the performance of our method by evaluating it using the following metrics: Attack Presentation Classification Error Rate (APCER) [20], Bona Fide Presentation Classification Error Rate (BPCER) [20], Average Classification Error Rate (ACER) = (APCER+BPCER)/2 [20] and Half Total Error Rate (HTER) = (False Acceptance Rate + False Rejection Rate)/2 [20].

4.2. Implementation Details

All data in the public datasets are provided in video format, so every frame is extracted from the videos and converted to an image file. For each frame, face regions are cropped and resized to 256 × 256, by using a face detector [50]. The input images are augmented to have six channels (RGB+HSV). Other than this, we do not use any data augmentation methods. The initialization method from [17] is used to initialize the weights of our network. For every dataset that we use, the number of spoof images is larger than the number of real images. Therefore, for every training epoch, we randomly sample negative images to set the ratio of positive and negative images as $1:1$. To train the network, we use a constant learning rate of 0.00005 with Adam optimizer [24]. Every experiment is conducted on a single NVIDIA Titan Xp GPU, and Tensorflow [1] framework is used to implement our work.

4.3. Ablation Study

We conduct ablation study to understand the effect of the proposed methods with three broad configurations:

(i) Baseline: The baseline model is designed to ablate the AFE module from BASN and is trained with only a softmax loss. Therefore, the feature map from the backbone network is not enriched with auxiliary feature maps. The remaining settings are the same as the proposed architec-

ture.
(ii) Baseline with auxiliary supervision: The models of this configuration are trained with unipartite auxiliary supervision. The AFE module of Model A and B is designed to have a single stream of depth feature extracting part and reflection feature extracting part, respectively. Therefore, feature fusion is performed only with an auxiliary feature map and the feature map from the backbone network. The remaining settings are the same as the proposed model.
(iii) Baseline with feature enrichment method: The models of this configuration are trained with bipartite auxiliary supervision with different feature enrichment methods. Instead of subtraction, Model C is designed to feature-wise add feature maps of depth and reflection, and Model D is the proposed BASN.
Each model is evaluated by following Oulu-NPU Protocol 1, and results are shown in Table 1.

Advantage of Bipartite Auxiliary Supervision. The baseline model with simple binary supervision shows the poorest performance with the highest ACER. Compared to the result of the binary supervision, both of the models with unipartite and bipartite supervision show better performances. Model D with bipartite supervision outperforms the models with unipartite supervision with a large margin. From this result, we can see the effectiveness of the bipartite auxiliary supervision.

Advantage of Proposed Feature Fusion Method. The result of Model C is worse than the result of Model A of unipartite supervision. However, the result of Model D with the proposed feature fusion method achieves lower ACER than the result of Model C by a large margin, as the subtraction allows a wider range of representation. The result indicates that the bipartite supervision is effective only if the proper feature fusion method is applied. With the proposed feature fusion methods, BASN shows notable improvements in performance and this demonstrates the effectiveness of our method.

4.4. Intra-Testing

Intra-testing is conducted on SiW and Oulu-NPU datasets. Different protocols are defined for each dataset and we strictly follow the defined protocols. Comparisons of our results with the existing methods are shown in Tables 2, 3.

Oulu-NPU. Our method outperforms the state-of-the-art results on Protocol 3 and 4 but shows higher ACER on Protocol 1 and 2. Our method shows a weakness of generalization ability when evaluated on Protocol 1, which evaluates generalization capabilities on illumination and background variations. For Protocol 2, our method achieves 0.5 percentage points (pp) higher ACER compared to the best result, and it is ranked as the 3^{rd} (Table 2). For Protocol 3, our method outperforms the state-of-the-art by a margin of 0.1 pp. For the most challenging Protocol 4, which evaluates a generalization ability of all possible variations in this dataset, our method outperforms the state-of-the-art by 0.8 pp.

Prot.	Method	APCER (%)	BPCER (%)	ACER (%)
1	CPqD [4]	2.9	10.8	6.9
	GRADIANT [4]	1.3	12.5	6.9
	MILHP [30]	8.3	**0.8**	4.6
	STASN [49]	**1.2**	2.5	1.9
	Auxiliary [32]	1.6	1.6	1.6
	FaceDe-S [21]	**1.2**	1.7	**1.5**
	Ours	1.5	5.8	3.6
2	MixedFASNet [4]	9.7	2.5	6.1
	MILHP [30]	5.6	5.3	5.4
	FaceDe-S [21]	4.2	4.4	4.3
	Auxiliary [32]	2.7	2.7	2.7
	GRADIANT [4]	3.1	1.9	2.5
	STASN [49]	4.2	**0.3**	**2.2**
	Ours	**2.4**	3.1	2.7
3	MixedFASNet [4]	5.3±6.7	7.8±5.5	6.5±4.6
	MILHP [30]	**1.5±1.2**	6.4±6.6	4.0±2.9
	GRADIANT [4]	2.6±3.9	5.0±5.3	3.8±2.4
	FaceDe-S [21]	4.0±1.8	3.8±1.2	3.6±1.6
	Auxiliary [32]	2.7±1.3	3.1±1.7	2.9±1.5
	STASN [49]	4.7±3.9	**0.9±1.2**	2.8±1.6
	Ours	1.8±1.1	3.6±3.5	**2.7±1.6**
4	MassyHNU [4]	35.8±35.3	8.3±4.1	22.1±17.6
	MILHP [30]	15.8±12.8	8.3±15.7	12.0±6.2
	GRADIANT [4]	5.0±4.5	15.0±7.1	10.0±5.0
	Auxiliary [32]	9.3±5.6	10.4±6.0	9.5±6.0
	STASN [49]	6.7±10.6	8.3±8.4	7.5±4.7
	FaceDe-S [21]	**1.2±6.3**	6.1±5.1	5.6±5.7
	Ours	6.4±8.6	**3.2±5.3**	**4.8±6.4**

Table 2. The intra-testing results of four protocols of Oulu-NPU dataset.

SiW. Our method achieves more noticeable improvements on SiW dataset. For every protocol defined in this dataset, our method achieves lower ACER than the state-of-the-art (Table 3). The ACER of our method is 63%, 57%, and 22% lower than the state-of-the-art for Protocol 1, 2 and 3, respectively.

4.5. Cross-Testing

The generalization capability of the proposed BASN is demonstrated by conducting cross-dataset evaluations. We use CASIA-FASD and Replay-Attack for the experiments and results are measured in HTER. Our approach shows better performance than the state-of-the-art on cross-dataset evaluation from CASIA-FASD to Replay-Attack with 4.0 pp lower HTER and achieves 1.5 pp higher HTER than

Prot.	Method	ACER (%)
1	Auxiliary [32]	3.58
	STASN [49]	1.00
	Ours	**0.37**
2	Auxiliary [32]	0.57±0.69
	STASN [49]	0.28±0.05
	Ours	**0.12±0.03**
3	STASN [49]	12.10±1.50
	Auxiliary [32]	8.31±3.80
	Ours	**6.45±1.80**

Table 3. The intra-testing results of three protocols of SiW dataset.

the state-of-the-art on evaluation from Replay-Attack to CASIA-FASD (Table 4). As an average score of both evaluations, our method achieves HTER of 26.75%, which is the lowest among all methods tested. This result is notable because cross-testing is usually more challenging than intra-testing. The cross-testing results of our method demonstrate that bipartite auxiliary supervision is effective to guide the network to capture faithful features that lead to improvements in generalization capability.

Method	Train	Test	Train	Test
	CASIA-FASD	Replay-Attack	Replay-Attack	CASIA-FASD
Motion [10]	50.2%		47.9%	
LBP-TOP [10]	49.7%		60.6%	
Motion-Mag [3]	50.1%		47.0%	
Spectral cubes [41]	34.4%		50.0%	
LBP [5]	47.0%		39.6%	
Color Texture[6]	30.3%		37.7%	
CNN [47]	48.5%		45.5%	
STASN [49]	31.5%		30.9%	
FaceDe-S [21]	28.5%		41.1%	
Auxiliary [32]	27.6%		**28.4%**	
Ours	**23.6%**		29.9%	

Table 4. The cross-testing results on CASIA-FASD vs. Replay-Attack.

5. Visualization

To better understand the behavior of the feature enrichment of BASN, we visualize distributions of the intermediate feature maps of BASN (Fig. 4) and a binary supervision model. Each mark of Fig. 4 denotes a test video of Oulu-NPU Protocol 1 and t-SNE [35] is used for visualization. By comparing the distribution of the output feature maps of the proposed feature enrichment method with the others, we can observe that the feature maps are more discriminately clustered. The ablation studies and visualization demonstrates that the proposed feature enrichment is effective and important.

Also, auxiliary maps estimated by BASN are shown in Fig. 5. The auxiliary maps are successfully estimated on images with various conditions, such as poses, genders, and expressions. Most of the estimated results are successful and selected results of failure cases are also presented.

6. Conclusions

In this paper, we present a new perspective of using auxiliary supervision for effective face anti-spoofing. Instead of providing faithful supervision of either live or spoof images, we suggest using bipartite auxiliary supervision that well represents characteristics of both live and spoof images. This bipartite auxiliary supervision helps to improve the generalization capabilities of the trained network. Also, we propose a bipartite auxiliary supervision network, which uses bipartite auxiliary supervision with a feature-enrichment method. The feature representation is enriched by fusing the extracted faithful features and the fused features help to detect presentation attacks with low error rates. We conduct several studies and experiments, and the results demonstrate the effectiveness of our work.

Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2014-0-00059, Development of Predictive Visual Intelligence Technology), and also supported by the MSIT, Korea, under the SW Starlab support program (IITP-2017-0-00897) supervised by IITP.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.

[2] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *International Joint Conference on Biometrics (IJCB)*, 2017.

[3] Samarth Bharadwaj, Tejas I Dhamecha, Mayank Vatsa, and Richa Singh. Computationally efficient face spoofing detection with motion magnification. In *CVPR Workshops*, 2013.

[4] Zinelabdine Boulkenafet, Jukka Komulainen, Zahid Akhtar, Azeddine Benlamoudi, Djamel Samai, Salah Eddine Bekhouche, Abdelkrim Ouafi, Fadi Dornaika, Abdelmalik Taleb-Ahmed, Le Qin, et al. A competition on generalized software-based face presentation attack detection in mobile scenarios. In *International Joint Conference on Biometrics (IJCB)*, 2017.

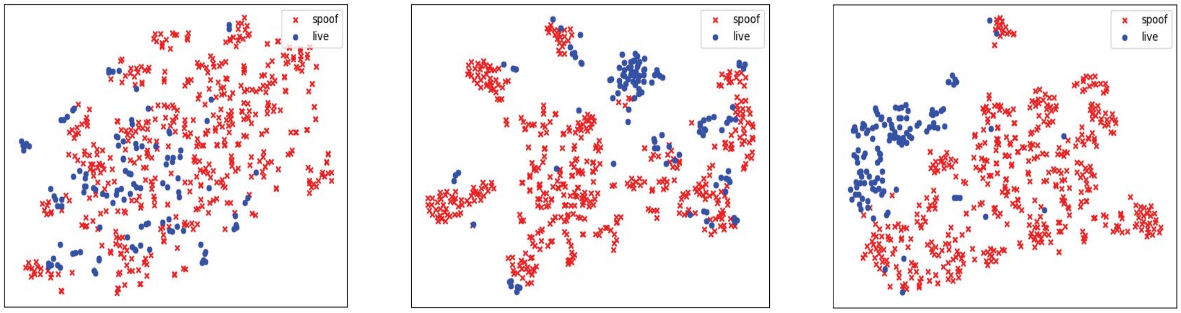

Figure 4. Distributions of intermediate feature maps from a model trained with only a softmax cross-entropy loss and from the proposed architecture. **Left**: distribution of feature maps from binary supervision model. **Middle**: distribution of depth feature maps of BASN. **Right**: distribution of the final output of the proposed feature enrichment method of BASN.

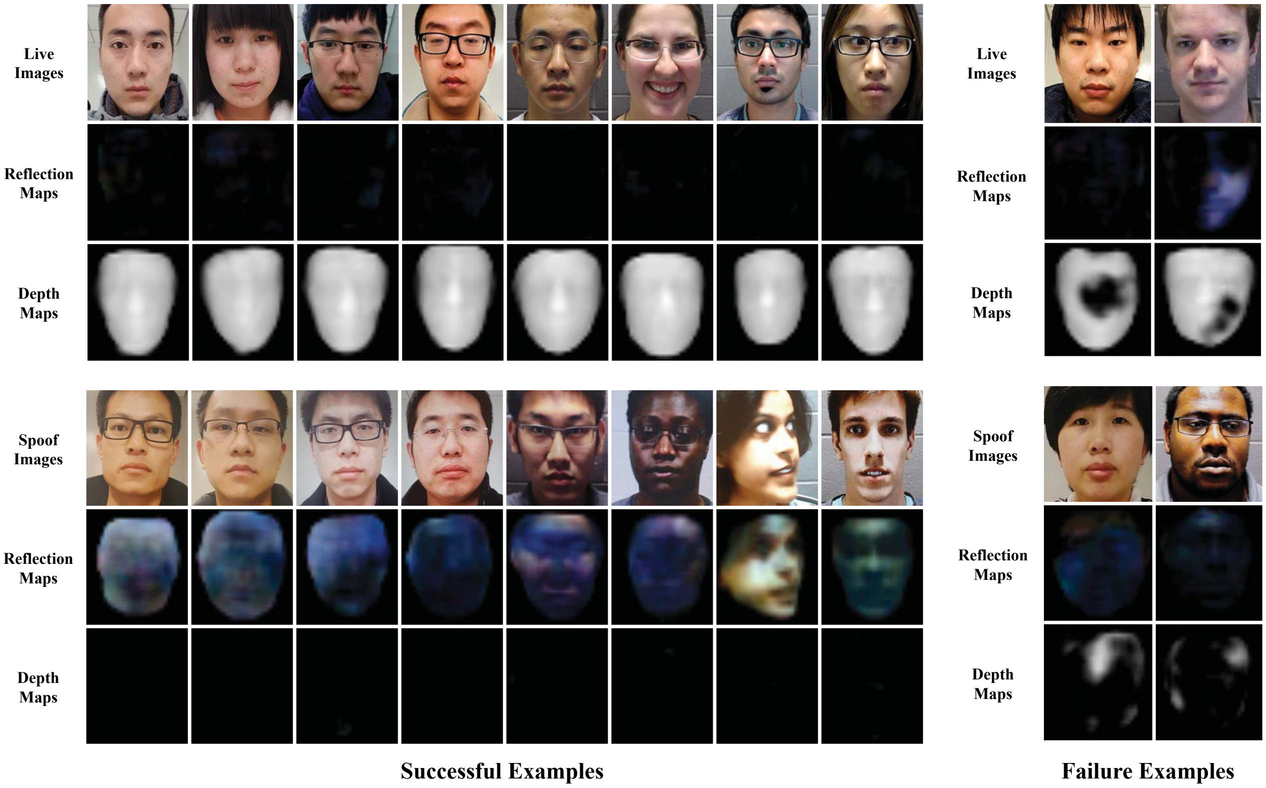

Figure 5. Visualization of input images and estimated auxiliary maps by BASN. The auxiliary maps are estimated by following Oulu-NPU protocol 4 and SiW protocol 1. For successful results, the first four columns are results of Oulu-NPU and the rest four are results of SiW. The first two columns of the successful results of each dataset are print-attacks and the rest two are video-attacks. For failure examples, the results of Oulu-NPU are presented in the first column and the results of SiW are presented on the other column.

[5] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *International Conference on Image Processing (ICIP)*, 2015.

[6] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8), 2016.

[7] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *International Conference on Automatic Face & Gesture Recognition (FG)*, 2017.

[8] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2012.

[9] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Lbp- top based countermeasure against face spoofing attacks. In *Asian Conference on Computer Vision (ACCV)*, 2012.

[10] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *International Conference on biometrics (ICB)*, 2013.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[12] Yueqi Duan, Jiwen Lu, and Jie Zhou. Uniformface: Learning deep equidistributed representation for face recognition. In *CVPR*, 2019.

[13] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018.

[14] Diogo Caetano Garcia and Ricardo L de Queiroz. Face-spoofing 2d-detection based on moiré-pattern analysis. *IEEE Transactions on Information Forensics and Security*, 10(4), 2015.

[15] Ross Girshick. Fast R-CNN. In *ICCV*, 2015.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[20] ISO/IEC JTC 1/SC 37 Biometrics. information technology biometric presentation attack detection part 1: Framework. (2016). `https://www.iso.org/obp/ui/iso`.

[21] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *ECCV*, 2018.

[22] Bong-Nam Kang, Yonghyun Kim, and Daijin Kim. Pairwise relational networks for face recognition. In *ECCV*, 2018.

[23] Wonjun Kim, Sungjoo Suh, and Jae-Joon Han. Face liveness detection from a single image via diffusion speed model. *IEEE Transactions on Image Processing*, 24(8), 2015.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[25] Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigun. Real-time face detection and motion analysis with application in liveness assessment. *IEEE Transactions on Information Forensics and Security*, 2(3), 2007.

[26] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Context based face anti-spoofing. In *International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[28] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2016.

[29] Lei Li, Xiaoyi Feng, Xiaoyue Jiang, Zhaoqiang Xia, and Abdenour Hadid. Face anti-spoofing via deep local binary patterns. In *International Conference on Image Processing (ICIP)*, 2017.

[30] Chen Lin, Zhouyingcheng Liao, Peng Zhou, Jianguo Hu, and Bingbing Ni. Live face verification with multiple instantialized local homographic parameterization. In *IJCAI*, 2018.

[31] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[32] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018.

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[34] Chaochao Lu and Xiaoou Tang. Surpassing human-level face verification performance on LFW with GaussianFace. In *AAAI*, 2015.

[35] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[36] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *International joint conference on Biometrics (IJCB)*, 2011.

[37] David Menotti, Giovani Chiachia, Allan Pinto, William Robson Schwartz, Helio Pedrini, Alexandre Xavier Falcao, and Anderson Rocha. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security*, 10(4), 2015.

[38] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *ICCV*, 2007.

[39] Keyurkumar Patel, Hu Han, and Anil K Jain. Cross-database face antispoofing with robust feature representation. In *Chinese Conference on Biometric Recognition (CCBR)*, 2016.

[40] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11(10), 2016.

[41] Allan Pinto, Helio Pedrini, William Robson Schwartz, and Anderson Rocha. Face spoofing detection through visual codebooks of spectral temporal cubes. *IEEE Transactions on Image Processing*, 24(12), 2015.

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[43] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, 2019.

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[45] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[46] Zhenqi Xu, Shan Li, and Weihong Deng. Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *Asian Conference on Pattern Recognition (ACPR)*, 2015.

[47] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv:1408.5601*, 2014.

[48] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In *International Conference on Biometrics (ICB)*, 2013.

[49] Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifeng Li, and Wei Liu. Face anti-spoofing: Model matters, so does data. In *CVPR*, 2019.

[50] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *ICCV*, 2017.

[51] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, 2018.

[52] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *International Conference on Biometrics (ICB)*, 2012.