A-DeepPixBis: Attentional Angular Margin for Face Anti-Spoofing

Md. Sourave Hossain*, Labiba Rupty*, Koushik Roy*, Md. Hasan*, Shirshajit Sengupta* and Nabeel Mohammed†
*Gaze Pte. Ltd., †North South University, Dhaka
{sourave, labiba, koushik, hasan, shirsho}@gaze.ai & nabeel.mohammed@northsouth.edu

***Abstract*—Face Anti Spoofing (FAS) systems are used to identify malicious spoofing attempts targeting face recognition systems using mediums such as video replay or printed papers. With increasing adoption of face recognition technology as a biometric authentication method, FAS techniques are gaining in importance. From a learning perspective, such systems pose a binary classification task. When implemented with Neural Network based solutions, it is common to use the binary cross entropy (BCE) function as the loss to optimize. In this study, we propose a variant of BCE that enforces a margin in angular space and incorporate it in training the DeepPixBis model [1]. In addition, we also present a method to incorporate such a loss for attentive pixel wise supervision applicable in a fully convolutional setting. Our proposed approach achieves competitive scores in both intra and inter-dataset testing on multiple benchmark datasets, consistently outperforming vanilla DeepPixBis. Interestingly, in the case of Protocol 4 of OULU-NPU, considered to be the hardest protocol, our proposed method achieves 5.22% ACER, which is only 0.22% higher than the current State of the Art without requiring any expensive Neural Architecture Search.**

Fig. 1. Difference between real and spoofed images. The first row in the figure showcases real photos and the second row shows spoofed faces.

I. Introduction

The convenient nature of face recognition systems makes it an attractive application for many settings. However, these systems can be very easily fooled by malicious presentation attacks (PA) making them quite susceptible. For reliable and secure deployment of face recognition systems, face anti-spoofing (FAS) systems play a very crucial role. Typically FAS systems provide protection against attacks using 3D masks, video replays, or printed media.

Previously most presentation attack detection (PAD) algorithms relied on handcrafted features that look for degradation of image quality while recapturing. Hand crafted feature based classic machine learning techniques such as [2], [3], [4], [5], [6] etc. rely heavily on these image quality degradation indications. With the wider availability of high quality cameras, such techniques are becoming less effective.

With these limitations in mind, recent literature has shown that several authors have opted for deep learning based approach for PAD. With the ability to represent complex features, Deep Convolutional Neural Networks (DCNN) have outperformed previously used handcrafted feature based techniques. Techniques such as [7], [1], [8], [9] uses DCNN to focus on deep semantic features to discriminate between spoofed faces and bona fide faces.

While techniques which use a video stream for PAD can be very effective, techniques which make frame-level decisions or decisions based on a single 2D image are very challenging but has the potential to be very efficient. Two recent studies from the literature for frame level PAD systems that leverage DCNNs are [1] and [7]. These studies have shown that DCNNs trained for frame-level PAD can be effective and generalize well as evaluated through inter-dataset evaluation. As we can see from figure 1, difference between spoofed and real images can be hard to spot. Sometimes they are inseparable even to human eyes.

Our proposed method is an extension of the DeepPixBis model proposed in [1]. We propose an angular margin based binary cross entropy loss function for both the classification layer as well as the fully convolutional pixel wise supervision head of the DeepPixBis model. Our results demonstrate that just by incorporating these loss adaptations, the modified model can outperform the original one [1] on 3 out of 4 protocols of the OULU-NPU dataset OULU-NPU dataset [10]. In particular, the proposed method improves ACER by almost 20% on Protocol 4 and is only 0.22% higher than the current state of the art (SOTA) [10] without using any expensive Neural Architecture Search methods. The contributions of this paper are:

- Inspired from the use of angular margin based loss functions for multi-class classification tasks, we propose an angular margin based binary cross-entropy loss function (A-BCE).
- While using angular binary cross-entropy is straight forward for the case where the final predictions involve a fully connected layer, it is not so obvious when making predictions using convolution layers. We

propose constraints to the convolution process which allows for easy incorporation of A-BCE.

- We introduce a learned attention tensor which is used to do a weighted summation of losses. This is used in the pixel-wise supervision head of the DeepPixBis model which is fully convolutional.

Also, the code for reproducible results along with all experiments can be found at https://github.com/gazeai/a-deeppixbis

II. Related Works

FAS techniques can be broadly categorized into video-based techniques, those that incorporate temporal information in the predictive model, and Frame level techniques that perform prediction on single 2D images. In this study, we are only concerned about frame level FAS techniques. Traditional Face Anti-Spoofing rely heavily on carefully extracted hand-crafted features. As discussed earlier, the literature is rich with studies that employ hand-crafted features for this task with more recent techniques using DCNNs. We briefly present a short over-view of both paradigms below.

A. Classic approaches

In this subsection we discuss FAS algorithms that are based on classical machine learning techniques. LBP [11], DoG [12], [3], HOG [2] etc. are some of the most popular techniques in classical FAS algorithms.

Boulkenafet *et al* [4] performs Fisher Vector [13], [14], [15] encoding on SURF (Speeded-up Robust Features) on different color spaces. The authors assume that there's a remarkable difference in color gamut between real images and spoofed images (print or replay attack). Spatially local luminance preserved by the algorithm varies because of the human eye's sensitivity. Obtained from using Wavelet transforms, the Haar box filters are used by the SURF descriptor. The CSURF (Colored Speeded-up Robust Feature) vectors are extracted from two color spaces (HSV and YCbCr). These vectors are concatenated and decorrelated with PCA after which they are processed into Fisher Vectors using the encoding method. These vectors are later fed into a classifier that classifies into real or attack classes. This rotation and translation invariant solution outperforms both most prior work in the same domain in inter-dataset testing.

Komulainen *et al*.[5] uses HOG [2] descriptors and Support Vector Machines (SVM) to perform the task of FAS. The main idea of the proposed method was to leverage the human context and scene information to determine spoofing. They also show that taking a close-up of fakes using two HOG [2] descriptors enhances the performance of this method. They also use alignment to further improve the performance. A head-and-shoulder detector was used to estimate the pose of the subject and align the face. This method has the best EER of 3.3% on CASIFA FAS dataset[16].

Another texture analysis based method was proposed by Li *et al*. [6] where the authors perform a 2D Fourier spectra of an image or a sequence of images to infer the authenticity of the image. Their work is based on two principles. Firstly, print faces have less high-frequency components. Secondly, live faces have subtle changes in the expression of multiple frames that printed faces don't have. In light of these two principles, the authors then calculated the HFD (high frequency descriptor) using both high and other frequency components. This method however fails when clearer and larger photos are shown, making this an easily spoof-able algorithm today.

Based on the work of Moriyama *et al.* [17], Sun *et al.* [18] proposes a blink based FAS method which utilizes Conditional Random Fields (CRF) and shows that liveness of a face can be determined through the activations of the human eye region. While other authors such as [17] work on the full resolution face image to determine liveness, [18] performs this only using the 24×24 crop of the eye regions. The author claims that CRFs are more robust due to the inherent ability to track long range sequences whereas prior methods are prone to error when bent or warped photos are presented to them. However, methods like these that highly depend on blinking will fail in scenarios where a replay attack is presented to the system.

Jukka *et al*. [11] proposes Uniform Patterns, a micro-texture analysis based solution that uses an extended version of Local Binary Pattern (LBP). These patterns are composed of a maximum of two bitwise transitions in the regions. Using this technique, through concatenation a histogram is generated which is fed to an SVM classifier to make the final prediction. The whole process consists of a few operations carried out in order. Initially, the detected face is cropped and normalized. Later, the face image is resized to a 64×64 image. After that, the LBP operator is applied on 3 by 3 overlapping regions of the original 64×64 image. Later, the generated histograms are concatenated and fed to the SVM to make the final prediction.

However, all of these methods lack in generalizability which is testified in the inter-dataset test results. Moreover, none of the aforementioned methods can extract deep features that Convolutional Neural Networks can.

B. Convolutional Neural Network based approaches

Introduced by Yang *et al*. [8], following an immense amount of research, Deep Convolutional Neural Network (DCNN) based approaches have become the new standard in spoof detection. For example, in last year's Multi-modal Face Anti-spoofing Attack Detection Challenge at CVPR, all of the 13 teams in the final round used a DCNN based method [19].

Yang *et al*. [8] proposes an improved generalized feature extraction method for FAS. The authors use the face alignment algorithm proposed by [20] and extract local binary features. These features are then processed through a modified layer of [21] and finally passed onto an SVM for classification.

Zero-Shot Face Anti-spoofing (ZSFA) can be defined as the instance where the model detects an unknown face attack. Focusing on print and replay attacks, [22], [23] rephrased this as an outlier detection problem to mitigate where the outlier being the unknown sample. Liu *et al*. [24] looks into 13 different types of spoof for the ZSFA problem and proposes a Deep Tree Network (DTN) to detect unknown spoof attacks. The method ensures that the unknown spoof types would always be traversed towards the leaf node of one of the 13 spoof types.

Li *et al*. [25] proposes an adversarial learning method for deep domain generalization. By adversarial feature learning,

this method proposes to align multiple source domains to an arbitrary prior distribution in order to learn the generalized feature space. Nonetheless, in order to be learned in the generalized feature space, there has to be data available from multiple source domains. This is solved by Shao *et al*. [26] who proposed a method to automatically search and learn the generalized distribution without utilizing any preceding distribution.

The anomaly detection method by [27] rephrased the FAS task into a deep metric learning problem. The authors proposed a "metric softmax" loss which uses the triplet focal loss[28] and guarantee an acceptable separability on the embedding space among the real and the attack classes.

Liu *et al*. [9] proposes a CNN-RNN based model for face depth estimation which also uses per pixel supervision and claims the unnecessity for further auxillary supervision. This model predicts remote photoplethysmography (rPPG) signals which is used as a feature for face liveliness. These features are further processed through a registration layer to form aligned feature maps that can be used for classification. In contrast, [29] utilises the noise informations of attack faces for a model which employs a combination of three different loss functions. However, the greatest model generalizability was shown by [29], [9] but with a caveat of requiring temporal informations which eventually leads to requiring higher number of frames. Furthermore, utilizing auxillary features like rPPG or depth eventually makes these models computationally expensive thus being difficult to deploy in remote devices like smartphones where the resources are limited.

The current state of the art for frame level face anti-spoofing is proposed by Yu *et al.* [7]. They propose a novel approach in the form of Central Difference Convolution (CDC) to the FAS domain. CDC is extremely capable of finding out detailed patterns via gradient and intensity. According to [7], a network built with CDC is capable of introducing more robust modeling in the FAS domain. The authors show their network built with CDC named Central Difference Convolutional Network (CDCN) outperforms the previous frame level SOTA model [29], [9]. They also do a Neural Architecture Search (NAS) to find an even more robust model called CDCN++. The CDCN++ beats CDCN in every protocol of OULU-NPU [10].

DeepPixBis [1] by George *et al.* proposes a framework that is efficient and accurate for frame level FAS. With pixel wise binary supervision [1] aims to simplify the requirement of having access to temporal information and complex depth map. They propose a fully convolutional network to generate a 14×14 score map. This 14×14 score map is used to perform the pixel wise binary supervision. Later this score map is flattened and fed into a fully connected layer paired with sigmoid activation to produce a binary output. The Binary cross-entropy loss is used in both the score map and binary output to guide the network. Our work is an extension of [1] and our contributions are centered around modifying the loss function to incorporate an angular margin with modifications to the model to support the required constraints.

III. Materials

In this section, we discuss the proposed method for face anti-spoofing for frame level spoof detection. The section starts with a description of all the datasets used for the experiments and their respective protocols. Following that we present the model architecture used in the experiments. Finally, we present our proposed angular binary cross-entropy (A-BCE), the constraints required to use it in a fully convolutional setting, and our attentive A-BCE for the pixel-wise supervision of DeepPix.

A. Datasets

This section discusses the datasets used for all the experiments we conducted. We primarily worked with 2 datasets: OULU-NPU [10] and Replay-Mobile [30]. For both datasets, we perform both intra-datset and inter-dataset experiments. Later in the results and experiments section, we show the metrics the proposed method achieved. A brief description of both the datasets is discussed below.

1) OULU-NPU: OULU-NPU [10] is the most recent dataset in the face anti-spoofing dataset family. It consists of 55 subjects and all the recordings were done using six phones and in three different environments (session). The session was introduced to better simulate real world situations. All the videos have a 1080p resolution. For print attacks, two printers were used to print HD images of the subjects. To enforce variation in the data, they used display devices as well. There were 3960 spoof videos (print and video replay combined) and 1980 bona fide videos. This makes it by far the most variant and definitely larger than MSU-MFSD [16] dataset. Also, there were 4 different protocols. The protocol names were set based on the level of difficulty. Following points show a brief description of OULU-NPU [10] protocols -

- Protocol I consists of a Test set which has different environments (session) than Dev and Train set.
- Protocol II ensures difference in instruments (smartphones) in the Test set as opposed to the Dev and Train set.
- Protocol III uses recording from different phones Train and Test set. This ensures models generalizability.
- Protocol IV is the hardest protocol of them all. It combines every constraints of the previous protocols. Also, with smaller subset of videos for training and evaluation.

2) Replay-Mobile: The replay-mobile dataset [30] has 1200 videos of 40 subjects in different conditions. A variety of illuminations were used to record these videos. Both well-lit and dimly-lit videos were recorded. Each subject has 10 videos in total in different background and illumination conditions. These videos are 10 seconds long and are recored in 720p resolution with an iPad mini 2 and an LG-g4 phone. There are mainly two types of attacks. They are -

- *Mattescreen*- Videos and photos are displayed on a HD monitor and then recorded off of it.
- *Print*- Faces were printed on A4 matte pages to demonstrate print attacks.

Protocol	Subset	Session	Phones	Users	Attacks created using	# real videos	# atttack videos	# all videos
Protocol I	Train	Session 1, 2	6 phones	1-20	Printer 1, 2; Display 1, 2	240	960	1200
	Dev	Session 1, 2	6 phones	21-35	Printer 1, 2; Display 1, 2	180	720	900
	Test	Session 3	6 phones	36-55	Printer 1, 2; Display 1, 2	120	480	600
Protocol II	Train	Session 1, 2, 3	6 phones	1-20	Printer 1; Display 1	360	720	1080
	Dev	Session 1, 2, 3	6 phones	21-35	Printer 1; Display 1	270	540	810
	Test	Session 1, 2, 3	6 phones	36-55	Printer 2; Display 2	360	720	1080
Protocol III	Train	Session 1, 2, 3	5 phones	1-20	Printer 1, 2; Display 1, 2	300	1200	1500
	Dev	Session 1, 2, 3	5 phones	21-35	Printer 1, 2; Display 1, 2	225	900	1125
	Test	Session 1, 2, 3	1 phones	36-55	Printer 1, 2; Display 1, 2	60	240	300
Protocol IV	Train	Session 1, 2	5 phones	1-20	Printer 1; Display 1	200	400	600
	Dev	Session 1, 2	5 phones	21-35	Printer 1; Display 1	150	300	450
	Test	Session 3	1 phones	36-55	Printer 2; Display 2	20	40	60

TABLE I. DIFFERNET PROTOCOLS OF OULU-NPU [10] AND THEIR CONFIGURATION

There are two protocols. Both of them are based on types of attack. It also has a grandtest protocol. Grandtest protocol is used to determine global performance. Throughout all our experiments, we used the grandtest protocol.

B. Metrics

For all our experiments, we used the ISO/IEC 30107-3 [31] metrics for evaluation. This is the current standard for FAS systems. Attack Presentation Classification Error Rate ($APCER$) was employed to calculate error rate among Presentation Attack Instances ($PAIs$) like print and video.

For the case of the model classifying real or bona fide faces as attack, Bona Fide Presentation Classification Error Rate ($BPCER$) was incorporated. Finally $ACER$ was used to evaluate the performance of the model as a FAS system. $ACER$ was computed as the mean of $APCER$ and $BPCER$.

$$ACER = \frac{APCER + BPCER}{2} \tag{1}$$

$$ACER = \frac{max_{forPAI=1,\ldots,X}(APCER_{PAI}) + BPCER}{2} \tag{2}$$

X here represents PA category type (print and/or video).

For inter dataset testing, Half Total Error Rate ($HTER$) was taken into consideration. The equation of $HTER$ is given below -

$$HTER = \frac{FAR + FRR}{2} \tag{3}$$

Where FAR and FRR is the False Acceptance Rate and False Rejection Rate of the system.

Equal Error Rate (EER) denotes the absolute difference between FAR and FRR. Following is the equation for EER -

$$EER = \|FAR - FRR\| \tag{4}$$

IV. METHODOLOGY

In this section, we discuss our proposed method for FAS. Our contribution is an angular margin loss based extension of [1]. Specifically we devise A-BCE, a form of the binary cross-entropy loss which leverages angular margins. To set the context of A-BCE, we first describe angular margin based losses used in multi-class classification. The following discussion about angular margin loss should naturally lead us to our proposed A-BCE, which is applicable for binary classification problem.

A. A brief look at Softmax loss

Before we dive into the method proposed by this paper for face anti-spoofing, let's take a look at softmax loss. Softmax is the most commonly used loss function for classification tasks. The mathematical formula for the softmax loss shown in equation 3:

$$L_{\text{softmax}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{5}$$

Here, x_{i} embedding vector of deep feature retrieved from the network for the i^{th} sample of class y_{i}. $W_j \in \mathbb{R}^m$ is the j_{th} column vector of the weight matrix $W \in \mathbb{R}^{m \times n}$. N and n are batch size and number of classes.

This loss function has been widely used for solving face recognition tasks in computer vision. However, as mentioned in [32], [33], [34], this loss function does not optimize properly to enforce higher intra-class similarity and inter-class dissimilarity. [33], [32], [34] tries to improve upon the already established softmax loss function by introducing a margin to the loss function.

For the sake of simplicity if we consider bias term b_j = 0, then $W_j^T \cdot x_i$ can be written as $\left\|W_j^T\right\| \|x_i\| \cos\theta_j$. Here θ denotes the angle between weight vector W_j^T and feature vector x_i for a specific sample. Now, if we normalize W_j^T and x_i to transform them into unit norm i.e. $\|W_j\|$ = 1 and $\|x_i\|$ = 1 by ℓ_2 normalization and re-scale $\|x_j\|$ to s, we are left with only $s\cos\theta_j$. So, 5 can be rewritten as:

$$L_{\text{ang}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{6}$$

As per authors of [34], if we introduce a hyper-parameter m then 6 becomes:

$$L_{\text{ang-margin}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_{\text{i}}}^{n} e^{s\cos\theta_j}} \tag{7}$$

Although softmax loss is mainly used for multi-class classification or used as categorical cross entropy, we can derive the same angular margin based loss function for binary cross-entropy as well. In the following sections we'll derive the proposed loss function from binary cross-entropy and discuss about network architecture and implementation detlails.

B. Angular binary cross-entropy Loss

Binary cross-entropy is a commonly used loss function for DCNN-based binary classification tasks and frequently found in relevant prior FAS studies. The popular equation for binary cross-entropy (BCE) is as follows:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N} p_i \log(s_i) + (1-p_i) log(1-s_i) \quad (8)$$

s_i in equation 8 is the predicted value and p_i is the ground truth value. N is the batch size. The output of the model, s_i, is usually calculated as the sigmoid activation of a transformation applied to an embedding vector (Equation 9).

$$s_i = \sigma(W^T \cdot x_i + b_i) \quad (9)$$

Here W and x_i denotes weights and incoming embedding/feature representation respectively and b_i is the bias term for i_{th} sample. σ is the sigmoid non-linearity which squeezes the output $W^T \cdot x_i$+b_i between 0 and 1. The equation of sigmoid function can be given as:

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (10)$$

Our work is an extension of [1]. In their paper [1] uses binary cross-entropy for both pixel wise and binary supervision. Binary cross-entropy paired with sigmoid non-linearity is the most widely used loss function for such tasks.

If we take a look at equations 8 and 9, we can see that it is possible to derive an angle based loss function from there. For the sake of simplicity, if we set bias b_i = 0 and ℓ_2 normalize both W_i and x_i so that $\|W_i\|$ = 1 and $\|x_i\|$ = 1, then :

$$\begin{aligned} s_i &= \sigma(W^T \cdot x_i) \\ &= \sigma(\|W\|\|x_i\| \cos\theta_i) \\ &= \sigma(\cos\theta_i) \end{aligned} \quad (11)$$

Where, θ is the angle between weight W and deep feature x_i. As we can see from equation 11, now the decision boundary depends on the angle between W and x_i rather than the value of these weights and deep features. All things considered, if we plug the new s_i into equation 8 then we get:

$$L_{\text{A-BCE}} = -\frac{1}{N}\sum_{i=1}^{N} p_i \log(\sigma(\cos\theta_i)) + (1-p_i) log(1-\sigma(\cos\theta_i)) \quad (12)$$

Now, to diverge the two classes $i.e.$ to increase intra-class similarity and inter-class dissimilarity, we also introduce a margin m. Specifically, we add the margin m with the angle θ to enforce a certain constraint in the angular space. After adding the margin equation 12 can be represented as equation 12

$$\begin{aligned} L_{\text{AM-BCE}} = -\frac{1}{N}\sum_{i=1}^{N} & p_i \log(\sigma(\cos(\theta_i + m))) \\ & + (1-p_i) log(1-\sigma(\cos\theta_i)) \end{aligned} \quad (13)$$

C. Proposed modifications

In this section we shall discuss about the proposed modification to [1]. First, we take a brief look at [1] and then we talk about modifications we propose.

DeepPixBis or deep pixel wise binary supervision [1] proposes a novel CNN based architecture for frame level FAS. Their main contribution is that they use pixel wise binary supervision to simplify the need for complex depth maps and how their proposed algorithm is deployment ready as it's a frame level FAS system with no requirement for temporal information. To do the pixel-wise binary supervision, they first obtain a 14×14 feature map M by convolving a 1×1 filter on the output of their chosen DenseNet [35] backbone. Thus, M is a 14×14 feature map with a single channel. Classification scores, S, are calculated at each point of M by simply taking the Sigmoid (Equation 10) of M. The pixel-wise supervision is then achieved by taking the mean loss $L_{pixel-wise-binary}$, as defined in Equation 14 where S is the sigmoid of M, h and w are indexing of the height and width of the feature map (in this case 14 for both) and y is the ground truth value for the image.

$$\begin{aligned} \mathcal{L}_{pixel-wise-binary} = -\frac{1}{14 \times 14}\sum_{h,w} & (y \log(S_{h,w}) + \\ & (1-y)\log(1-S_{h,w}) \end{aligned} \quad (14)$$

In addition, they also calculate an image-level classification score by transforming M into an embedding vector and passing it through a fully connected layer with a sigmoid activation to obtain a single score g on the entire image. The binary supervision loss L_{binary} obtained by calculating BCE (Equation 8) using g.

The total loss L is then simply calculated as a weighted sum of $L_{pixel-wise-binary}$ and L_{binary} as shown in Equation 15.

$$\mathcal{L} = \lambda \mathcal{L}_{pixel-wise-binary} + (1-\lambda)\mathcal{L}_{binary} \quad (15)$$

Here, λ is a hyper-parameter which was empirically set to 0.7 and for all experimental results reported in this paper we use the same value.

Inspired by the reported better discriminatory ability and the success of angular margin based loss functions in facial recognition tasks [34], [32], [33], we add binary angular margin loss to both parts of equation 15 using equation 13.

It is straightforward to replace $\mathcal{L}_{binary}$ with an equivalent version using Equation 13. The additional steps involve ensuring that while calculating the loss the entries in the weight matrix and the embedding/feature vectors are unit norm in addition to setting the bias to zero. To replace $L_{pixel-wise-binary}$ with an angular margin equivalent we needed to modify M, the feature map obtained from the backbone to $\hat{M}$ where $\hat{M}$ is obtained by normalizing M at each position along the channel axis. Then the actual convolution operation is performed on $\hat{M}$ where the 1×1 filter was also normalised along the channel axis. The bias of this convolution layer was set to 0. These little modifications ensured that the resulting 14×14 single channel feature map consisted of scores indicating the $cos(\theta_{a,b})$ values

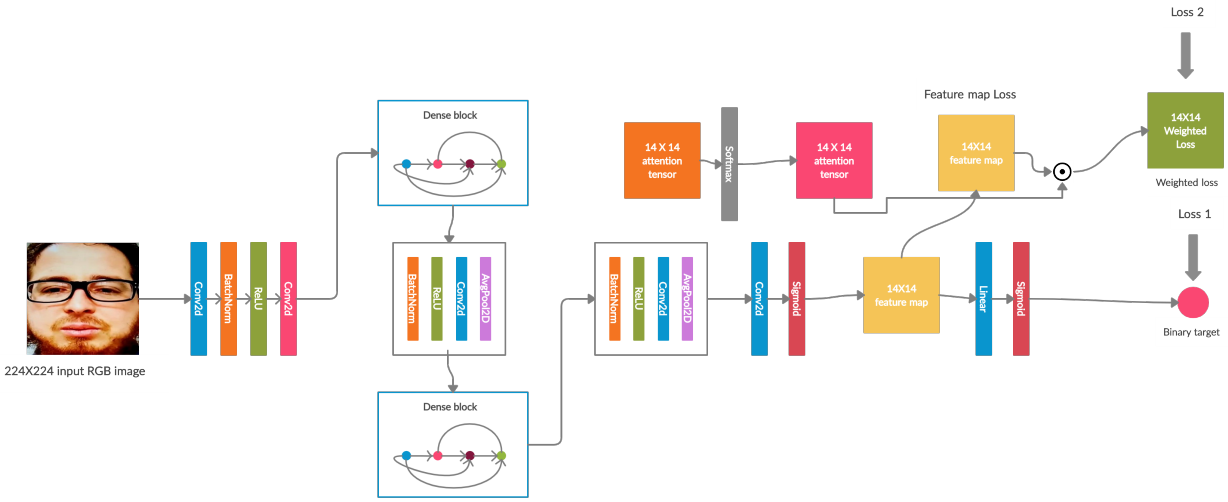

Fig. 2. Represents the whole framework. The backbone is composed of first 8 blocks of densenet161[35]. The proposed modifications take place at the 14×14 attention tensor end and loss 1 and loss 2

where the $\theta_{a,b}$ is the angle between the filter and the channel-wise vector at position (a, b) in the feature map $\hat{M}$.

In the original DeepPixBis implementation, these pixel-wise loss values were averaged to obtained $L_{pixel-wise-binary}$, as shown in Equation 14. We modify this scheme to do a weighted sum of the pixel-wise angular margin loss values, instead of taking a simple mean. Let $l_{a,b}$ be the angular loss calculated from position (a, b) of $\hat{M}$. We use a trainable tensor A of size $(14, 14)$ and calculate $\hat{A}$ by taking the softmax of A as shown in Equation 16.

$$\hat{A}_{ab} = \frac{e^{A_{ab}}}{\sum_{xy} e^{A_{xy}}} \tag{16}$$

Now, using $\hat{A}$ we can calculate the weighted sum of the angular pixel-wise supervision loss as show in Equation 17.

$$\mathcal{L}_{ang-pixel-wise-binary} = \sum_{ij} \hat{A}_{ij} \times l_{ij} \tag{17}$$

D. Modified network architecture

For the backbone network we used DenseNet161 [35] with imagenet pretrained weights. We initialize the weights with imagenet weights, but we don't freeze the layers. First eight layers of DenseNet161 [35] were used and the rest were discarded. The figure 2 shows the complete network architecture with all the modifications discussed previously.

The output of the first eight layer of DenseNet [35] is $384 \times 14 \times 14$ (channel first settings). Then we add a 1×1 Conv2D layer with sigmoid layer to get the 14×14 score map. To this end, the network architecture stays the same as [1]. From here onwards, we make the following changes to the network architecture:

- We add a 14×14 attention tensor A initialized with ones. We take the softmax of this attention tensor.
- We calculate the angular margin loss of the 14×14 score map S.
- We perform a Hadamard product of A and S to get the weighted loss $\mathcal{L}_{pixel-wise-binary}$. After that we plug the value of weighted loss value to equation 15 to get the total loss $\mathcal{L}$.

E. Implementation details

First we cropped faces from the videos using RetinaFace [36]. Unlike [1] we didn't align the faces. To deal with class imbalance in the dataset, majority class was under sampled. We used random sampling to balance out the dataset. For augmentation we did random horizontal flip, random color jitters. Learning rate of $1 \times e^{-3}$ with an weight decay of $1 \times e^{-5}$ was employed. We initialized the backbone with imagenet weights and the attention tensor with 1. Adam optimizer was used to optimize the network. We trained the network for 10 epochs and mini batch size of 64 was used. Throughout all out experiments the value of λ was set to 0.7. Following the work of [34] the value of m was set to 0.5.

For evaluation we used $APCER$, $BPCER$ and $ACER$ as measurement metrics. Outputs of both binary target and binary feature map was used to evaluate the network. In case of the 14×14 feature map, mean of the output tensor was used to determine the class label.

V. Experiments and Results

In this section we will discuss about the experiments conducted with the proposed method and their results. We'll compare out results with [1], current frame level SOTA [7] and also other notable algorithms. Metrics of all the algorithms reported in this section were taken from [1] and [7].

This section is divided into two subsections. In the first subsection, we show our experiment results from intra-dataset testing. We did all our experiments on Replay-Mobile [30] and OULU-NPU [10] datasets. In the later section, we show results of inter-dataset experiments.

A. Intra-dataset Testing

In this section we will discuss about intra-datset performance of the proposed method. For intra-datset testing we used both OULU-NPU [10] and Replay-mobile [30] datasets.

For OULU-NPU [10] we followed the protocols mentioned in the paper. In the case of the Replay-mobile [30] dataset, evaluation was done on the "$grandtest$" protocol. Metrics of other algorithms reported in this section were taken from [1] and [7]. Table II showcases intra-dataset testing results on OULU-NPU [10] dataset.

Protocol	Model	APCER(%)	BPCER(%)	ACER(%)
1	CPqD	2.9	10.8	6.9
	GRADIANT	1.3	12.5	6.9
	FAS-BAS	1.6	1.6	1.6
	IQM-SVM	19.17	30.83	25.0
	LBP-SVM	12.92	51.67	32.29
	CDCN	**0.4**	1.7	1.0
	CDCN++	**0.4**	**0.0**	**0.2**
	DeepPixBiS	0.83	**0.0**	0.42
	A-DeepPixBis(binary output)	0.83	0.58	0.7
	A-DeepPixBis(feature map)	1.19	0.31	0.75
2	MixedFASNet	9.7	2.5	6.1
	FAS-BAS	2.7	2.7	2.7
	GRADIANT	3.1	1.9	2.5
	IQM-SVM	12.5	16.94	14.72
	LBP-SVM	30	20.28	25.14
	CDCN	**1.5**	1.4	1.5
	CDCN++	1.8	**0.8**	**1.3**
	DeepPixBiS	11.39	0.56	5.97
	A-DeepPixBis(binary output)	4.07	1.4	2.74
	A-DeepPixBis(feature map)	4.35	1.29	2.82
3	MixedFASNet	5.3 ± 6.7	7.8 ± 5.5	6.5 ± 4.6
	GRADIANT	2.6 ± 3.9	5.0 ± 5.3	3.8 ± 2.4
	FAS-BAS	2.7 ± 1.3	3.1 ± 1.7	2.9 ± 1.5
	IQM-SVM	21.94 ± 9.99	21.95 ± 16.79	21.95 ± 8.09
	LBP-SVM	28.5 ± 23.05	23.33 ± 17.98	25.92 ± 11.25
	CDCN	2.4 ± 1.3	2.2 ± 2.0	2.3 ± 1.4
	CDCN++	**1.7 ± 1.5**	**2.0 ± 1.2**	**1.8 ± 0.7**
	DeepPixBiS	11.67 ± 19.57	10.56 ± 14.06	11.11 ± 9.4
	A-DeepPixBis(binary output)	2.78 ± 3.47	11.16 ± 16.45	6.97 ± 7.57
4	MassyHNU	35.8 ± 35.3	8.3 ± 4.1	22.1 ± 17.6
	GRADIANT	5.0 ± 4.5	15.0 ± 7.1	10.0 ± 5.0
	FAS-BAS	9.3 ± 5.6	10.4 ± 6.0	9.5 ± 6.0
	IQM-SVM	34.17 ± 25.89	39.17 ± 23.35	36.67 ± 12.13
	LBP-SVM	41.67 ± 27.03	55.0 ± 21.21	48.33 ± 6.07
	CDCN	4.6 ± 4.6	9.2 ± 8.0	6.9 ± 2.9
	CDCN++	4.2 ± 3.4	**5.8 ± 4.9**	**5.0 ± 2.9**
	DeepPixBiS	36.67 ± 29.67	13.33 ± 16.75	$25.0 \pm 12.$
	A-DeepPixBis(binary output)	**3.86 ± 4.04**	6.56 ± 7.88	5.22 ± 2.96

TABLE II. PERFORMANCE COMPARISON OF PROPOSED METHOD WITH OTHER ALGORITHMS ON OULU-NPU [10] INTRA-DATSET TESTING

As we can see our proposed method constantly achieve better $ACER$ than [1] (Except Protocol 1). Not only that, we also achieve very competitive $ACER$ on all 4 protocols of OULU-NPU [10]. Especially on Protocol 4 which is the hardest protocol of OULU-NPU [10], we achieve an $ACER$ of 5.22 ± 2.96 which is on par with the current SOTA for protocol 4 CDCN and CDCN++ [7]. In Protocol 4 our model achieves 0.22% higher $ACER$ than [7]. We achieve that just by making a simple change to the loss function. No expensive NAS or architecture design was necessary.

We also achieve good metrics in the "$grandtest$" protcol of Replay-Mobile [30] dataset. As we can see from Table III, our proposed method achieve EER and $HTER$ of 0% which is on par with current SOTA algorithms.

Model	EER(%)	HTER(%)
IQM-SVM	1.2	3.9
LBP-SVM	6.2	12.1
DeepPixBiS	**0.0**	**0.0**
A-DeepPixBis(binary output)	**0.0**	**0.0**
A-DeepPixBis(feature map)	**0.0**	**0.0**

TABLE III. PERFORMANCE COMPARISON OF PROPOSED METHOD WITH OTHER ALGORITHMS ON REPLAY-MOBILE "$grandtest$" PROTOCOL [30]

B. Inter-datset Testing

We also performed inter-dataset testing to validate our proposed algorithm. The purpose of inter-dataset testing is to demonstrate the generalization ability of a PAD algorithm. For inter-dataset testing we trained a model with OULU-NPU [10] and tested on Replay-Mobile (RM) [30]. Table IV shows the proposed method outperforms [1] in inter-dataset testing as well.

	Trained on OULU		Trained on RM	
Model	tested on OULU	tested on RM	tested on OULU	tested on RM
IQM-SVM	24.6	31.6	**3.9**	42.3
LBP-SVM	32.2	35.0	12.1	43.6
DeepPixBiS	**0.4**	12.4	22.7	**0.0**
A-DeepPixBis(feature map)	0.7	**9.35**	25.57	**0.0**

TABLE IV. INTER-DATASET TESTING OF OULU-NPU("$protocol1$") AND REPLY-MOBILE "$grandtest$" PROTOCOL. REPORTED METRICS IN THE TABLE REPRESENTS HTER VALUES IN PERCENTAGE(%).

As we can see from Table IV, although the proposed approach generalizes well in OULU-RM inter-dataset testing, it scores higher $APCER$ on the RM-OULU test than vanilla DeepPixBis [1]. We assume this is due to lack of variance in Replay-Mobile [30] dataset as opposed to OULU-NPU [10] dataset. With more data and variance in illumination and environment, our proposed method generalizes well enough.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we propose a novel approach to reconstruct the binary cross-entropy as a loss function for FAS systems. In particular, we propose an angular margin based binary cross entropy loss function and demonstrate methods to calculate such a loss for attentive pixel-wise supervision in a fully convolutional setting. We incorporate our modifications to the DeepPixBis model [1] and show that our proposed changes lead to superior results in most cases.The results also show that just by changing the loss function we obtain error rates that are very competitive with the current frame level SOTA [7]. Inter-data set testing demonstrates that our method generalizes well across data sets collected independently. One question that might arishe is why use A-BCE over vanilla BCE. From the literature [33], [32], [34] it's very obvious that the discriminatory power of angular margin loss over its vanilla counterpart gives A-BCE an edge . In future, we plan to incorporate our angular margin based binary cross entropy loss function to other popular FAS architectures like CDCN and CDCN++ and compare the results.

REFERENCES

[1] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *International Conference on Biometrics*, no. CONF, 2019.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, 2005.

[3] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *European Conference on Computer Vision*, pp. 504–517, Springer, 2010.

[4] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and fisher vector encoding," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 141–145, 2016.

[5] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–8, IEEE, 2013.

[6] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of fourier spectra," in *Biometric Technology for Human Identification*, vol. 5404, pp. 296–303, International Society for Optics and Photonics, 2004.

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

[8] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *arXiv preprint arXiv:1408.5601*, 2014.

[9] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 389–398, 2018.

[10] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "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, IEEE, 2017.

[11] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *2011 international joint conference on Biometrics (IJCB)*, pp. 1–7, IEEE, 2011.

[12] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," in *2011 18th IEEE International Conference on Image Processing*, pp. 3557–3560, IEEE, 2011.

[13] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods.," in *BMVC*, vol. 2, p. 8, 2011.

[14] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild.," in *BMVC*, vol. 2, p. 4, 2013.

[15] J. Sánchez and F. Perronnin, "High-dimensional signature compression for large-scale image classification," in *CVPR 2011*, pp. 1665–1672, IEEE, 2011.

[16] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face anti-spoofing database with diverse attacks," in *2012 5th IAPR international conference on Biometrics (ICB)*, pp. 26–31, IEEE, 2012.

[17] T. Moriyama, T. Kanade, J. F. Cohn, J. Xiao, Z. Ambadar, J. Gao, and H. Imamura, "Automatic recognition of eye blinking in spontaneously occurring behavior," in *Object recognition supported by user interaction for service robots*, vol. 4, pp. 78–81, IEEE, 2002.

[18] L. Sun, G. Pan, Z. Wu, and S. Lao, "Blinking-based live face detection using conditional random fields," in *International Conference on Biometrics*, pp. 252–260, Springer, 2007.

[19] A. Liu, J. Wan, S. Escalera, H. Jair Escalante, Z. Tan, Q. Yuan, K. Wang, C. Lin, G. Guo, I. Guyon, *et al.*, "Multi-modal face anti-spoofing attack detection challenge at cvpr2019," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[20] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1685–1692, 2014.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[22] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13868–13882, 2017.

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

[24] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4680–4689, 2019.

[25] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018.

[26] R. Shao, X. Lan, J. Li, and P. C. Yuen, "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.

[27] D. Pérez-Cabo, D. Jiménez-Cabello, A. Costa-Pazo, and R. J. López-Sastre, "Deep anomaly detection for generalized face anti-spoofing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[29] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 290–306, 2018.

[30] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–7, IEEE, 2016.

[31] ISO/IEC, "Information technology — Biometric presentation attack detection — Part 3: Testing and reporting," standard, International Organization for Standardization, Geneva, CH, sep 2017.

[32] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

[33] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

[34] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

[35] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[36] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *arXiv preprint arXiv:1905.00641*, 2019.