Abstract

Face biometrics have attracted significant attention in many security-based applications. The presentation attack (PA) or face spoofing is a cybercriminal attempt to gain illegitimate access to a victim’s device using photos, videos, or 3D artificial masks of the victim’s face. Various deep learning approaches can tackle particular PA attacks when tested on standard datasets. However, these methods fail to generalize to complex environments or unseen datasets. We propose a new Multi-Teacher Single-Student (MTSS) visual Transformer with a multi-level attention design to improve the generalizability of face spoofing detection. Then, a novel Multi-Level Attention Module with a DropBlock (MAMD) is designed to strengthen discriminative features while dropping irrelevant spatial features to avoid overfitting. Finally, these rich convolutional feature sets are combined and fed into the MTSS network for face spoofing training. With this MAMD module, our method survives well under small training datasets with poorly lighted conditions. Experimental results demonstrate the superiority of our method when compared with several anti-spoofing methods on four datasets (CASIA-MFSD, Replay-Attack, MSU-MFSD, and OULU-NPU). Furthermore, our model can run on Jetson TX2 up to 80 FPS for real-world applications.

1 Introduction

Face biometrics have wide applications in facility security and smart devices for identifying authentic accesses from users [12]. Biometric face recognition provides a passive, seamless,
Figure 1: Our network architecture consists of a three-branch visual Transformer with CNN backbones and a newly proposed Multi-Level Attention Module with DropBlock (MAMD) design. Each CNN block is connected to a maximum pooling layer. Figure 2 and Figure 3 will provide additional details on the visual Transformer and MAMD, respectively.

and touchless solution that is preferred over traditional methods such as badges or manual password entries in these applications. As a result, it often serves as the first line of defense in information security for convenience and effectiveness. Although face biometric systems are widely used, they are in fact vulnerable to spoofing and attacks. In biometrics, liveness detection\(^1\) is a computer’s ability to determine that it is interfacing with a physically present human being and not an inanimate spoof artifact or injected video/data. The presentation attacks (PA) or face spoofing is a cybercriminal attempt to gain illegitimate access to a victim device using photos, videos, or 3D artificial masks of the victim’s face. Face spoofing attacks can invade face biometric systems by showing a printed photo, recorded video, or 3D artificial mask in front of the camera to try granting access. To safeguard facial biometric systems, it is important to develop a reliable method that can identify the subtle difference between real and spoofing faces to filter out malicious attacks effectively.

In recent years, advancements in convolutional neural networks (CNNs) have achieved excellent performance in object detection, classification, and related analysis tasks. Most recent Face Anti-Spoofing (FAS) research works adopt CNN-based methods \([26, 27, 38]\) to extract deep features. In \([26]\), a CNN-RNN model estimates the face depth and heart pulse rPPG signals with sequence-wise supervision to distinguish live vs. spoof faces. In \([38]\), a CNN model learns to discriminative deep dynamic textures for 3D mask face anti-spoofing. In \([27]\), a Central Difference Convolutional Network (CDCN) captures the detailed intrinsic patterns by aggregating both intensity and gradient information for FAS. This model often fails to work in complex environments when trained with insufficient data.

An alternative to improving FAS detection is to incorporate additional sensors such as depth camera \([\text{[4]}]\), infrared (IR) irradiation \([\text{[42]}]\), or thermal cameras. However, liveness detection of human faces using additional modalities is not a common practice in biometrics applications considering the cost, portability, and availability issues.

In this paper, we develop an RGB image-based FAS method that can reliably detect face spoofing without relying on additional imaging modality or hardware. Our pipeline (see Figure 1) consists of (1) a visual Transformer feature extractor (§ 3.1) followed by conv layers and (2) a Multi-Level Attention Module with DropBlock (MAMD) (§ 3.2) followed by addi-

\(^1\)From https://liveness.com/.
tional conv and FC layers. Our three-branch Transformer can better extract YCbCr features that are robust against lighting variations. In addition, the MAMD design can strengthen the learning of discriminative features while dropping irrelevant spatial features to address overfitting issues during training better. This pipeline is trained as the teacher and student networks in § 3.3 following a Multi-Teacher Single-Student (MTSS) training paradigm (Figure 5). The resulting student model can be deployed on edge or mobile devices running web-based applications. Compared to existing state-of-the-art face anti-spoofing approaches on public datasets (in Section 4), the proposed method features the following design improvements:

- The proposed MAMD scheme can better address overfitting during training by strengthening the discriminative features while dropping irrelevant spatial features.
- The MAMD module can capture enough detailed textures and survives well under a small training set and poor lighting conditions.
- The MTSS learning framework can effectively train a lightweight student model, which can run at 80FPS on a NVIDIA Jetson TX2 board for real-world applications.
- Extensive evaluations and comparisons against 11 methods on 4 public datasets (CASIA-MFSD [59], Replay-Attack [9], MSU-MFSD [51], and OULU-NPU [7]) show that our approach outperforms the competing methods.

2 Related Work

**Face Anti-Spoofing.** Biometric anti-spoofing methods [12] can be categorized into: (1) optical sensor based [21], (2) computer vision based [48], and (3) hybrid (sensor and computer vision) methods [30]. Real-world applications of face anti-spoofing technology usually require high detection accuracy with low computational resources. As the discriminative capacity usually limits image-based methods, more and more works combine the image-based method with other approaches including remote photoelectric volume scanning [40], time series [32], and texture detection [28] for anti-spoofing. However, due to the convenience and deployment cost considerations most face biometric systems use an RGB camera as the only imagery device. Therefore, RGB-based anti-spoofing methods remain a major research trend.

**Traditional** face spoofing detection methods [1, 24, 32] mainly use handcrafted image-based features can operate on single or multiple image frames. Single-frame anti-spoofing methods leverage features such as LBP [1, 10], SIFT [35], SURF [6], BSIF [31], HOG [23] and adopt common classifiers such as LDA and SVM for decision making. Fourier spectrum in the HSV or YCbCr color space is typically used for spoofing detection. Multi-frame anti-spoofing methods can leverage motion and dynamic cues, such as eye blinking [32, 41], lip and mouth movement [24]. However, in practice, spontaneous facial motion is very subtle and thus hard to capture using traditional handcrafted image features.

**CNN-based** (Convolutional Neural Networks) approaches [19, 24, 34] have become the mainstream for face liveness and spoofing detection due to their robust feature extraction and discriminative capabilities, where features are extracted using convolutional operations, and the network weights and parameters learned end-to-end. Typically a few fully connected layers based on softmax are performed at the end of the CNN to predict spoofing classification. CNNs are also used for spoofing detection methods based on depths [13, 58].
Additional cues such as multi-frames [33] or rPPG [16, 25] signals can capture dynamic information and enrich the characterization of time series for spoofing detection. Although CNN based anti-spoofing methods are widely used, they often encounter overfitting and scalability issues [39] when tested across datasets. Although the incorporation of multiple sensing modalities (such as depth [33, 57, 58]) can improve performance, they may not meet the real-world requirements in practical applications.

**Teacher/Student Optimization** [17] is a simple yet extremely useful network training scheme in deep learning. Semi-supervised knowledge transfer is leveraged during training to improve the overall accuracy of a neural network. A teacher network is trained first, then the prediction knowledge of the teacher will be used to guide the training of a student network, where the teacher/student network can consist of the same architecture. This teacher/student paradigm was originally intended for knowledge distilling [17] to train a larger teacher network and compress the learned model into a smaller student network [37]. There are also methods designed to work with deeper or wider models [8, 20], where the pre-training weights are initialized to predict the narrower or shallower models. Several extensions are proposed to improve the supervision of the teacher network [44, 45]. In [36, 46], multiple teacher networks are used to guide a single student network. In [54], the teacher/student concept is used to gradually adjusting the network optimization, where the next learning status is supervised according to the results from the current training iteration.

Although knowledge distilling [17] and model compression [37] have been successfully applied in various domains, there is no mature method applied to face anti-spoofing detection tasks. Therefore, we propose a multi-teacher and single-student approach that allows teachers to specialize in different learning domains, computes learning errors for both teachers and students, and distills knowledge to students through teachers. In addition, we propose an attention mechanism, similar to the diffusion and erosion approach, combined with the concept of multiscale to extract the true and false features of faces more efficiently.

### 3 Method

Figure 1 overviews our network pipeline, which consists of a Transformer with CNN backbones followed by a Multi-Level Attention Module with DropBlock (MAMD) with additional conv and FC layers. Section 3.1 describes our three-branch YCbCr Transformer feature extractor. Section 3.2 describes the MAMD design to generate better multi-level features for FAS. The proposed network pipeline is used in the teacher/student networks following a multi-teacher training scheme, which will be explained in Section 3.3.
Figure 3: The proposed MAMD module refines the Transformer extracted feature maps via a High-Low Attention Block (HLAB, detailed in Figure 4) after the DropBlock [14].

### 3.1 Visual Transformer

The Transformer architecture with attention mechanism [49] was initially developed for sequence-to-sequence [43] translation and understanding in Natural Language Processing (NLP). Various Transformer mechanisms have been widely developed for sentiment analysis, machine translation, speech recognition, and dialogue robots [47]. Recently, in computer vision, visual Transformer is increasingly become as important as RNN and CNN due to its self-attention mechanism, which helps models to focus on only certain parts of the input and reason more effectively [15]. Unlike CNN, a visual Transformer can retain the original characteristics of the image while modeling temporal sequence relations. Computer vision tasks that traditionally adopt CNN and RNN [29] can now be better handled using attention-based visual Transformers [11].

Figure 2 shows our proposed three-branch visual Transformer feature extractor. This module takes an input RGB image of $256 \times 256$. The three color channels are converted into YCbCr color space, where the Y, Cb, Cr channels are rearranged into a 1-D feature vector and fed into the visual Transformer for feature extraction. Positional information is added to the feature map through the Position Embedding of the Transformer. The multi-head attention mechanism of the Transformer can generate richly attended feature maps, which are finally converted and resized to match the input size.

### 3.2 Multi-Level Attention Module with DropBlock (MAMD)

CNNs can often extract major characteristics while ignoring the subtle features that are important for spoofing detection [55]. To this end, we propose a Multi-Level Attention Module with DropBlock (MAMD) that can solve two major problems of overfitting and feature pooling for FAS across domains of large variations. MAMD can effectively refine and fuse discriminative features via multi-level convolutions (i.e. low-, medium-, and high-level convolutions, respectively), in reducing incorrect attentions occurred in typical CNN features.

**Addressing overfitting.** The Transformer feature maps $F$ from Section 3.1 are fed into multi-level (\{low, medium, high\}) convolutional layers to extracted discriminative features as in Figure 3. Let $\odot$ denote the Hadamard matrix operator [52] that takes two input matrices of identical dimensions and produces a resulting matrix of the same dimension. We design a novel High-Low Attention Block (HLAB) as in Figure 4, where min-pooling and max-
pooling are performed to aggregate spatial difference information on the feature map. The HLAB is crucial for MAMD to retain useful features for anti-spoofing detection.

**HLAB to highlight attention on discriminant features.** In grayscale morphological analysis, the *erosion* and *dilation* operators can be extended to the use of *max* and *min* filters to extract texture features. Motivated from the morphological analysis, we adopt the *min-pooling* and *max-pooling* followed by convolutions with different mask sizes to extract discriminant features for FAS. As shown in Figure 4, the min-pooling (denoted by *MinP*) and max-pooling (denoted by *MaxP*) are performed on the feature map to aggregate texture information on the feature maps $F$. After concatenation, different low-, medium-, and high-level texture feature maps are further extracted via convolution masks $C_i$ with different mask sizes $9 \times 9$, $7 \times 7$, and $5 \times 5$, respectively. Next, an activation function ‘Leaky ReLU’ is adopted to calculate spatial attentions [52, 58] on feature maps to endue the network with discriminative features. Let $i \in \{\text{low}, \text{medium}, \text{high}\}$ corresponding to mask sizes $\{9 \times 9, 7 \times 7, 5 \times 5\}$, respectively. At the $i^{th}$ level, after the HLAB module, the feature map $HLAB_i(F)$ with attentions can be generated as follows:

$$HLAB_i(F) = \text{LeakyReLU}(C_i(\text{MinP}(F) + \text{MaxP}(F))),$$ (1)

where $+$ in this equation denotes a concatenation operator.

**DropBlock to enhance the robustness of feature map.** Dropout is a common regularization technique by randomly omitting (dropping) activation units. Inspired by DropBlock [14], after the HLAB modules, pixels in a contiguous group of a feature map are dropped together to enhance its robustness. For example, let $\text{DropBlock}(F)$ denote a randomly generated block to drop groups of activation in $F$. In FAS, each group is composed of activation units in a continuous region to delete certain semantic information (such as hair or glasses) relatively efficiently. Then, $HLAB_i(F)$ can be converted to a new feature map:

$$F_{MAMD}^i = \text{DropBlock}(F) \odot HLAB_i(F),$$ (2)

where $F_{MAMD}^i$ denotes the final feature map obtained by the MAMD block, then all the feature maps $\{F_{MAMD}^i\}$ are fed into an FC (fully-connected) classifier for FAS identification.

### 3.3 MTSS: Multi-Teacher Single-Student

We adopt the MTSS learning scheme [36] initially developed for spoken dialog systems to train our anti-spoofing network pipeline. The *multi-teacher* scheme can better circumvent the complex multi-domain state representation in capturing subtle but important features
Figure 5: The proposed MTSS learning paradigm can learn spoofing attacks on data-1 and data-2 from two teacher-1 and teacher-2 respectively, and use the learned knowledge to train the student network to achieve better spoofing detection performance on both dataset split.

4 Experiments and Results

Experiments of the proposed method are performed on four mainstream public face anti-spoofing datasets described in Section 4.1. Section 4.2 provides our training and implementation details. Section 4.3 describes the evaluation metrics and Section 4.4 discusses observations from experiments and highlights effectiveness and benefits of our method.

4.1 Datasets

The OULU-NPU dataset [1] is a high-resolution dataset consisting of 4950 real and fake face videos. The Presentation Attacks (PA) are created with 2D images. Four different types of Protocols (Protocol 1 to 4 in Table 2) are used to evaluate the effectiveness of any face anti-spoofing detection method. The three datasets CASIA-MFSD [23], Replay-Attack [9],
Table 1: Cross-type test AUC(%) evaluation results of the models on the three data sets of CASIA-MFSD [59], Replay-Attack [9] and MSU-MFSD [51].

<table>
<thead>
<tr>
<th>Method</th>
<th>CASIA-MFSD</th>
<th>Replay-Attack</th>
<th>MSU-MFSD</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Video</td>
<td>Cut Photo</td>
<td>Wrapped Photo</td>
<td>Printed Photo</td>
</tr>
<tr>
<td>OC-SVM RBF+BSIF</td>
<td>70.74</td>
<td>60.73</td>
<td>95.9</td>
<td>84.03</td>
</tr>
<tr>
<td>SVM RBF+LBP</td>
<td>91.94</td>
<td>91.7</td>
<td>84.47</td>
<td>99.08</td>
</tr>
<tr>
<td>NN+LBP</td>
<td>94.16</td>
<td>88.39</td>
<td>79.85</td>
<td>99.75</td>
</tr>
<tr>
<td>DTN [27]</td>
<td>90</td>
<td>97.3</td>
<td>97.5</td>
<td>99.9</td>
</tr>
<tr>
<td>CDCN [57]</td>
<td>98.48</td>
<td>99.9</td>
<td>99.8</td>
<td>100</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>96.96</td>
<td>98.81</td>
<td>98.06</td>
<td>99.99</td>
</tr>
</tbody>
</table>

and MSU-MFSD [51] we used are in low image resolution. The PA types for CASIA-MFSD [59] include Video, Cut Photo, Wrapped Photo. The PA types for Replay-Attack [9] include Video, Digital Photo, Printed Photo. Furthermore the PA types for MSU-MFSD [51] include Printed Photo, HR Video, Mobile Video as all the datasets mentioned above consist only of videos for both training and testing. Each frame is resized to 256 × 256 with face alignment. When testing, only a single image is given as the input.

### 4.2 Training

During training, all the pixel values were normalized to within the range [0,1]. To train the MTSS module, the learning difficulty increases according to iterations by increasing drop probability from 0.1 to 0.5 in the DropBlock [14] mentioned in Section 3.2.

All experiments are performed on the platform with RTX TITAN GPU for training. The initial learning rate is set to 1e^{-4}, and weight decay is 2e^{-5} in the Adam Optimizer. The MTSS is trained for 200 epochs with a batch size of 64. The learning rate is dropped by 0.1 after every 20 epochs.

### 4.3 Evaluation Metrics

For experiments in the CASIA-MFSD [59], Replay-Attack [9], and MSU-MFSD [51] datasets, we split the data and test it out according to the datasets’ formats. We use the standard Area-Under-Curve (AUC) of the Receiver Operating Characteristic (ROC) curve to evaluate the face spoofing detection performance.

For experiments performed on the OULU-NPU [8] dataset, we use the recent ISO/IEC 30107-3 biometric standard indicators [4] for evaluation. We use Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error rate (BPCER) for real face (known from groudtruth) and face spoofing identification, respectively. The Average Classification Error Rate (ACER) is the error rate when classifying a real attack correctly. Given the true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), APCER, BPCER, and ACER are defined as:

\[
APCER = \frac{FP}{TN + FP}, \quad BPCER = \frac{FN}{FN + TP}, \quad \text{and,} \quad ACER = \frac{APCER + BPCER}{2}. \quad (3)
\]
4.4 Results and Discussions

Extensive evaluations and comparisons with 11 methods, namely OC-SVMRBF+BSIF [8], SVM RBF+LBP [3], NN+LBP [23], DTN [12], CDCN [33] (in Table 1), and GRADIENT [1], STASN [28], Auxiliary [22], FaceDs [11], FAS-TD [36], DeepPixBiS [4], and CDCN (in Table 2), on 4 public data sets, namely CASIA-MFSD [32], Replay-Attack [1], MSU-MFSD [31], and OULU-NPU [7]. Tables 1 and 2 shows comparison results, where scores of the comparing methods come from their original papers.

Considering the resistance of the models in consideration to unknown attack types, we choose the Oulu-NPU database [7] for our model verification and strictly follow the four Protocols of OULU-NPU [7]. In addition, considering the diversity of attacks, we choose CASIA-MFSD [32], Replay-Attack [1], and MSU-MFSD [31] databases to verify the stability of the model against diverse attacks and follow the database standards for verification. When testing, we use the attack to present the APCER classification error rate. For the face anti-spoofing, we use BPCER, ACER, and the ROC-AUC as the evaluation criteria. In Tables 1, our method gets 100% accuracy in the Printed Photo category for the Replay-Attack [7] dataset. For the MSU-MFSD dataset [31], most State-of-The-Art (SoTA) methods fail to detect attacks in the Printed Photo category with lower accuracy. Our analysis shows that most SoTA methods require high-definition (HD) images with sharp features to work well. They cannot detect attacks of small-sized, poor-lighted images, and thus are not suitable for webcam-based FAS. In addition, these SoTA methods require a lot of training samples to train the classifiers to achieve good accuracy for FAS. The numbers of training samples are 40248, 88023, and 31699, respectively. Tables 1 shows that our method performs better on relatively smaller datasets.

Table 2 shows the comparisons among different SoTA methods on the OULU-NPU [7] dataset. There are four protocols for performance comparisons. Protocol 1 and Protocol 2 are similar but with different numbers of training samples; that is, 1200 and 1080, respectively. In the two protocols, the phones used for training and testing are the same. In Protocol 2, our method outperforms than other SoTA methods for all metrics. Protocol 3 and Protocol 4 are similar, they both use the different phones for training and testing. But their number of training samples is very different; that is, 1500 and 600, respectively. Thus, Protocol 4 is more challenging than the others since the dataset is small with more challenges. In Protocol 3, CDCN [22] performs better than our method; however we note that our method were trained for only 200 epoches without data augmentation. For the most challenging ‘Protocol 4’, our method outperforms other SoTA methods, even though limited amount of training samples are provided. Protocol 4 is close to the real environments especially for webcam-
based applications, where only fewer training samples are available. Clearly, our method survives well even though few samples with bad lighted conditions are provided.

### 4.5 Ablation Experiment

We perform an ablation study on OULU-NPU Protocol 4 and MSU-MFSD to show the effectiveness of our proposed modules.

As shown in Table 3, we report APCER, BPCER, and ACE of protocol 4 on OULU-NPU [1], and AUC of Printed Photo, HR Video, Mobile Video on the MSU-MFSD [51]. We compare the performance of the baseline and each module (and their combinations), it is clear that each proposed module and their combinations introduce performance gain on the two datasets, which shows the effectiveness of our proposed method.

Specifically, the individual module (MAMD and HLAB) witness significant performance gain compared to the baseline CNN architecture on most (5 out of 6) of the settings (APCER, BPCER, ACER, HR and Printer ). The combination of MAMD and HLAB can further improve the performance, and the best performance is achieved with our whole model (MTSS + MAMD + HLAB).

<table>
<thead>
<tr>
<th></th>
<th>OULU-NPU Protocol 4</th>
<th>MSU-MFSD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>APCER ↓</td>
<td>BPCER ↓</td>
</tr>
<tr>
<td>Baseline</td>
<td>20.1</td>
<td>10.1</td>
</tr>
<tr>
<td>MAMD</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>HLAB</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>MAMD+HLAB</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>MTSS+MAMD+HLAB</td>
<td>6.6</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Table 3: Ablation study on OuLu Protocol-4 and MSU-MFSD.

### 5 Conclusion

This paper proposes a transformer-based architecture that transforms input images in YCbCr color space. A multi-Level Attention Module with a DropBlock (MAMD) is applied to produce the rich feature for face anti-spoofing detection. We have also proposed a Multi-Teacher Single-Student (MTSS) based network, which can effectively works in different challenging scenarios to identify the face anti-spoofing in different challenging datasets. The extensive experiments on various public datasets can achieve promising results on different image qualities and presentation attacks. Moreover, merging our approach with a novel Multi-Teacher Single-Student (MTSS) model can improve the overall performance compared to the State-of-The-Art and run on embedded devices with Jetson TX2 with 80 FPS for real-world applications. The teacher-student learning mechanism improve the applicability of our method for real-world applications, in particularly for cases when only very few samples with poor lighting conditions are available.

**Future work.** For all performance evaluations, data augmentation is not adopted in our method to get better accuracy. For Protocol 3, the accuracy of our method can be further improved. In addition, the stochastic weight averaging [18] will be adopted in our scheme for further performance improvement.
References


