Multi-Teacher Single-Student Visual Transformer with Multi-Level Attention for Face Spoofing Detection

Yao-Hui Huang[1]
10857023@email.ntou.edu.tw
Jun-Wei Hsieh[2]
jwhsieh@nctu.edu.tw
Ming-Ching Chang[3]
mchang2@albany.edu
Lipeng Ke[4]
lipengke@buffalo.edu
Siwei Lyu[4]
siweilyu@buffalo.edu
Arpita Samanta Santra[1]
santraarpita83@gmail.com

[1] National Taiwan Ocean University, Keelung, Taiwan
[2] National Yang Ming Chiao Tung University, Tainan, Taiwan
[3] University at Albany, State University of New York, Albany, New York, 12222, USA
[4] University at Buffalo, State University of New York, Buffalo, New York, 14260, USA

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,

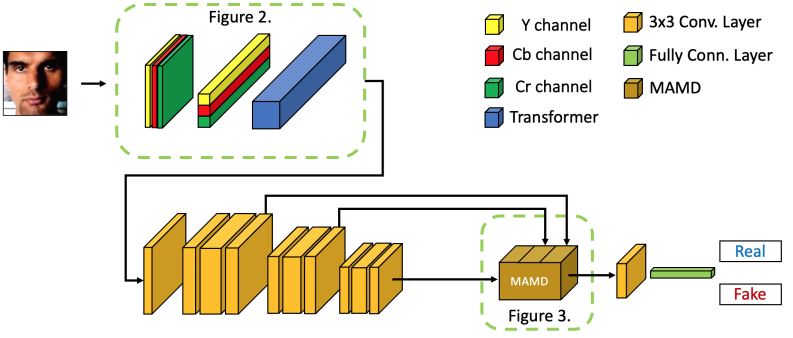

Figure 1: Our network architecture consists of a three-branch visual Transformer with CNN backbones and a newly proposed Multi-Level Attention Module with DropBlock (MAMD) design. Each CNN block is connected to a maximum pooling layer. Figure 2 and Figure 3 will provide additional details on the visual 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 [1], infrared (IR) irradiation [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 [22, 32, 41] mainly use handcrafted image-based features can operate on single or multiple image frames. *Single-frame* anti-spoofing methods leverage features such as LBP [8, 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 [22]. 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 [13, 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 [33, 57, 58].

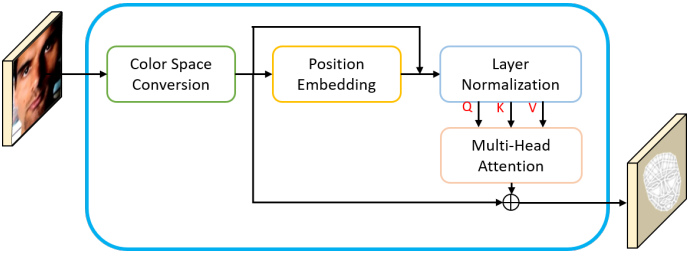

Figure 2: The input face image is first converted to YCbCr and reshaped into a sequence to be fed into the Transformer feature extractor.

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.

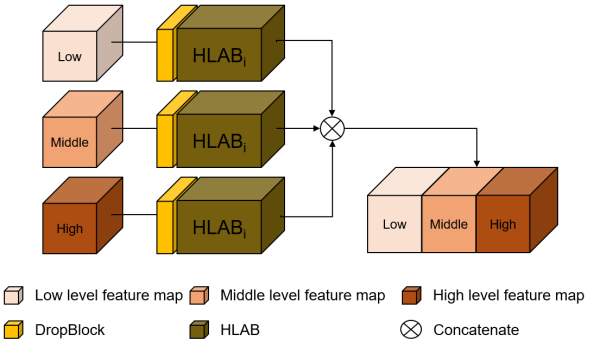

Figure 3: The proposed MAMD module refines the Transformer extracted feature maps via a High-Low Attention Block (HLAB, detailed in Figure 4) after the DropBlock [14].

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 256×256. 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-

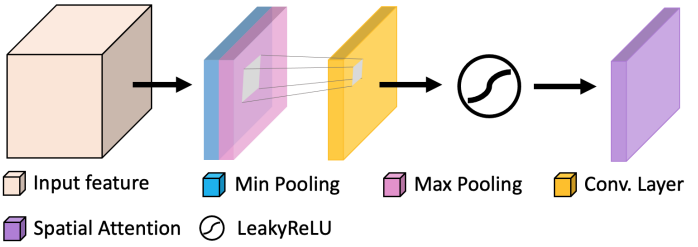

Figure 4: The High and Low Attention Block (HLAB) to aggregate discriminative spatial features for spatial attention learning.

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 9×9, 7×7, and 5×5, 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 $i \in \{low, medium, high\}$ 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:

$$HLAB_i(F) = LeakyRELU(C_i(MinP(F) + MaxP(F))), \tag{1}$$

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 $DropBlock(F)$ 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:

$$F^i_{MAMD} = DropBlock(F) \odot HLAB_i(F), \tag{2}$$

where F^i_{MAMD} denotes the final feature map obtained by the MAMD block, then all the feature maps $\{F^i_{MAMD}\}$ 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

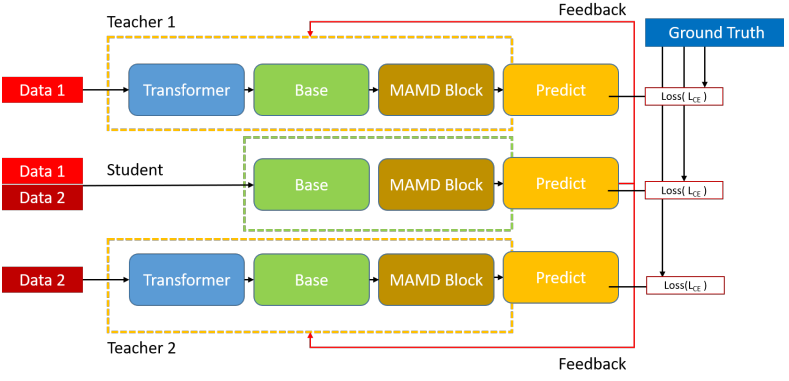

Figure 5: The proposed MTSS learning paradigm can learn spoofing attacks on data-1 and data-2 from two teacher-1 and teacher-2 respectively , and use the learned knowledge to train the student network to achieve better spoofing detection performance on both dataset split.

for learning a face spoofing detector that is scalable across variations and data domains, e.g., different types of mobile phones. Instead, each teacher network can focus on specific domain knowledge based on a precisely extracted state representation.

Figure 5 depicts our two-teacher, single-student training paradigm. We first divide the batch data set into two parts:(i) spoofing attacks using *printed faces* (Data 1) and (ii) spoofing attacks using a *cell phone display* (Data 2). We then assign two teacher networks to learn the two scenarios separately, where the learned knowledge will be used to train the student network on the whole data batch. In case when the student network found questions regarding the learning (*,e.g.*, the test answer is inconsistent), the question will be handed over to the teacher network to re-learn. In other words, such inconsistent information (the arrow with 'red line' in Figure 5) will be noted in the next iteration of the teacher network training. The teacher network will either reward or penalize the student network based on the learning experiences, which will reflect in the loss of the student network. In case when the teacher network cannot learn any further, we will penalize both the teacher and student networks all-together, which will reflect in the loss of both networks.

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 [7] 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 [59], Replay-Attack [9],

Method	CASIA-MFSD [59]			Replay-Attack [9]			MSU-MFSD [51]			Overall
	Video	Cut Photo	Wrapped Photo	Video	Digital Photo	Printed Photo	Printed Photo	HR Video	Mobile Video	
OC-SVMRBF+BSIF [7]	70.74	60.73	95.9	84.03	88.14	73.66	64.81	87.44	74.69	78.68 ± 11.74
SVM RBF+LBP [7]	91.94	91.7	84.47	99.08	98.17	87.28	47.68	99.5	97.61	88.55 ± 16.25
NN+LBP [63]	94.16	88.39	79.85	99.75	95.17	78.86	50.57	99.93	93.54	86.69 ± 16.25
DTN [27]	90	97.3	97.5	99.9	99.9	99.6	81.6	99.9	97.5	95.9 ± 6.2
CDCN [57]	98.48	99.9	99.8	100	99.43	99.92	70.82	100	99.99	96.48 ± 9.64
Ours	96.96	98.81	98.06	99.99	**99.9**	**100**	**96.7**	97.65	99.58	$\mathbf{98.68 \pm 2.35}$

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].

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 creases 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 [7] 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:

$$APCER = \frac{FP}{TN+FP}, \quad BPCER = \frac{FN}{FN+TP}, \quad \text{and,} \quad ACER = \frac{APCER+BPCER}{2}. \quad (3)$$

Method	Protocol 1			Protocol 2			Protocol 3			Protocol 4		
	APCER	BPCER	ACER	APCER	BPCER	ACER	APCER	BPCER	ACER	APCER	BPCER	ACER
GRADIANT [8]	1.3	12.5	6.9	3.1	1.9	2.5	2.6±3.9	5.0±5.3	3.8±2.4	5.0±4.5	15.0±7.1	10.0±5.0
STASN [56]	1.2	2.5	1.9	4.2	0.3	2.2	4.7±3.9	0.9±1.2	2.8±1.6	6.7±10.6	8.3±8.4	7.5±4.7
Auxiliary [26]	1.6	1.6	1.6	2.7	2.7	2.7	2.7±1.3	3.1±1.7	2.9±1.5	9.3±5.6	10.4±6.0	9.5±6.0
FaceDs [19]	1.2	1.7	1.5	4.2	4.40	4.3	4.0±1.8	3.8±1.2	3.6±1.6	1.2±6.3	6.1±5.1	5.6±5.7
FAS-TD [50]	2.5	0.0	1.3	1.7	2.0	1.9	5.9±1.9	5.9±3.0	5.9±1.0	14.2±8.7	4.2±3.8	9.2±3.4
DeepPixBiS [13]	0.8	0.0	0.4	11.4	0.6	6	11.7±19.6	10.6±14.1	11.1±9.4	36.7±29.7	13.3±14.1	25.0±12.7
CDCN [57]	0.4	1.7	1.0	1.5	1.4	1.45	2.4±1.3	2.2±2.0	2.3±1.4	4.6±4.6	9.2±8.0	6.9±2.9
Ours	2	1	1.5	**1.4**	**0.3**	**0.85**	**2.1±1.3**	**0.5±0.4**	**1.3±0.8**	6.6±3.3	**2.4±2.8**	**4.5±2.2**

Table 2: Evaluation results in APCER, BPCER, and ACER scores (%) comparing the 7 anti-spoofing methods and ours on the OULU-NPU [7] dataset.

4.4 Results and Discussions

Extensive evaluations and comparisons with **11** methods, namely OC-SVMRBF+BSIF [7], SVM RBF+LBP [7], NN+LBP [53], DTN [27], CDCN [57] (in Table 1), and GRADIANT [8], STASN [56], Auxiliary [26], FaceDs [19], FAS-TD [50], DeepPixBiS [13], and CDCN (in Table 2), on **4** public data sets, namely CASIA-MFSD [59], Replay-Attack [9], MSU-MFSD [51], 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 [59], Replay-Attack [9], and MSU-MFSD [51] 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 [9] dataset. For the MSU-MFSD dataset [51], 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 for the CASIA-MFSD [59], Replay-Attack [9], and MSU-MFSD [51] datasets 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 [57] 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-

	OULU-NPU Protocol 4 [7]			MSU-MFSD [51]		
	APCER ↓	BPCER ↓	ACER ↓	HR ↑	Mobile ↑	Printer ↑
Baseline	20.1	10.1	15.1	90.43	94.86	93.71
MAMD	12	6	9	96.48	95.09	94.70
HLAB	13	6	10	95.74	94.69	95.18
MAMD+HLAB	9	5	7	96.03	96.14	97.45
MTSS+MAMD+HLAB	6.6	2.4	4.5	97.06	97.65	99.58

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

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 [7], 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).

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 obtain better accuracy. For Protocol 3, the accuracy of our method can be further improved. In addition, stochastic weight averaging [18] can be adopted to further improve performance in our scheme.

References

[1] Ghazel Albakri and Sharifa Alghowinem. The effectiveness of depth data in liveness face authentication using 3d sensor cameras. *Sensors*, 19(8):1928, 2019.

[2] Shervin Rahimzadeh Arashloo, Josef Kittler, and William Christmas. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE access*, 5:13868–13882, 2017.

[3] Samarth Bharadwaj, Tejas I Dhamecha, Mayank Vatsa, and Richa Singh. Computationally efficient face spoofing detection with motion magnification. In *CVPR Workshop*, pages 105–110, 2013.

[4] ISO/IEC JTC1 SC37 Biometrics. Information technology–biometric presentation attack detection – Part 3: testing and reporting, 2017.

[5] 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 *IJCB*, pages 688–696. IEEE, 2017.

[6] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 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 *FG*, pages 612–618. IEEE, 2017.

[8] Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2Net: Accelerating learning via knowledge transfer. *arXiv:1511.05641*, 2015.

[9] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, pages 1–7. IEEE, 2012.

[10] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. LBP-TOP based countermeasure against face spoofing attacks. In *ACCV*, pages 121–132. Springer, 2012.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

[12] Javier Galbally, Sébastien Marcel, and Julian Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014.

[13] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

[14] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. DropBlock: A regularization method for convolutional networks. *arXiv:1810.12890*, 2018.

[15] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao. A survey on visual transformer. In *arXiv:2012.12556*, 2020.

[16] Javier Hernandez-Ortega, Julian Fierrez, Aythami Morales, and Pedro Tome. Time analysis of pulse-based face anti-spoofing in visible and NIR. In *CVPR Workshop*, pages 544–552, 2018.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

[18] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[19] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *ECCV*, pages 290–306, 2018.

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016.

[21] Sooyeon Kim, Yuseok Ban, and Sangyoun Lee. Face liveness detection using a light field camera. *Sensors*, 14(12):22471–22499, 2014.

[22] Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigun. Real-time face detection and motion analysis with application in "liveness" assessment. *TIFS*, 2 (3):548–558, 2007.

[23] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Context based face anti-spoofing. In *BTAS*, pages 1–8. IEEE, 2013.

[24] 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 *IPTA*, pages 1–6. IEEE, 2016.

[25] Bofan Lin, Xiaobai Li, Zitong Yu, and Guoying Zhao. Face liveness detection by rPPG features and contextual patch-based CNN. In *ICBEA*, pages 61–68, 2019.

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

[27] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *CVPR*, pages 4680–4689, 2019.

[28] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using texture and local shape analysis. *IET biometrics*, 1(1):3–10, 2012.

[29] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv:2103.03841*, 2021.

[30] Dat Tien Nguyen, Tuyen Danh Pham, Min Beom Lee, and Kang Ryoung Park. Visible-light camera sensor-based presentation attack detection for face recognition by combining spatial and temporal information. *Sensors*, 19(2):410, 2019.

[31] Olegs Nikisins, Amir Mohammadi, André Anjos, and Sébastien Marcel. On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In *ICB*, pages 75–81. IEEE, 2018.

[32] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *ICCV*, pages 1–8. IEEE, 2007.

[33] Aleksandr Parkin and Oleg Grinchuk. Recognizing multi-modal face spoofing with face recognition networks. In *CVPR Workshop*, 2019.

[34] Keyurkumar Patel, Hu Han, and Anil K Jain. Cross-database face antispoofing with robust feature representation. In *CCBR*, pages 611–619. Springer, 2016.

[35] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *TIFS*, 11(10):2268–2283, 2016.

[36] Shuke Peng, Feng Ji, Zehao Lin, Shaobo Cui, Haiqing Chen, and Yin Zhang. MTSS: Learn from multiple domain teachers and become a multi-domain dialogue expert. In *AAAI*, pages 8608–8615, 2020.

[37] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. *arXiv:1412.6550*, 2014.

[38] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Joint discriminative learning of deep dynamic textures for 3D mask face anti-spoofing. *TIFS*, 14(4):923–938, 2018.

[39] Amit K Shukla, Manvendra Janmaijaya, Ajith Abraham, and Pranab K Muhuri. Engineering applications of artificial intelligence: A bibliometric analysis of 30 years (1988–2018). *Engineering Applications of Artificial Intelligence*, 85:517–532, 2019.

[40] Talha Ahmad Siddiqui, Samarth Bharadwaj, Tejas I Dhamecha, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Face anti-spoofing with multifeature videolet aggregation. In *ICPR*, pages 1035–1040. IEEE, 2016.

[41] Lin Sun, Gang Pan, Zhaohui Wu, and Shihong Lao. Blinking-based live face detection using conditional random fields. In *ICB*, pages 252–260. Springer, 2007.

[42] Shengtao Sun, Yue Tian, Ying Tang, and Ben Wu. Anti-spoofing face recognition using infrared structure light. In *Frontiers in Optics*, pages FW1F–3, 2020.

[43] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv:1409.3215*, 2014.

[44] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[45] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *AAAI*, volume 31, 2017.

[46] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv:1703.01780*, 2017.

[47] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. In *arXiv:2009.06732*, 2020.

[48] Santosh Tirunagari, Norman Poh, David Windridge, Aamo Iorliam, Nik Suki, and Anthony TS Ho. Detection of face spoofing using visual dynamics. *TIFS*, 10(4):762–777, 2015.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[50] Zezheng Wang, Chenxu Zhao, Yunxiao Qin, Qiusheng Zhou, Guojun Qi, Jun Wan, and Zhen Lei. Exploiting temporal and depth information for multi-frame face anti-spoofing. *arXiv:1811.05118*, 2018.

[51] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *TIFS*, 10(4):746–761, 2015.

[52] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, pages 3–19, 2018.

[53] Fei Xiong and Wael AbdAlmageed. Unknown presentation attack detection with face rgb images. In *BTAS*, pages 1–9. IEEE, 2018.

[54] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan Yuille. Knowledge distillation in generations: More tolerant teachers educate better students. *arXiv:1805.05551*, 2018.

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

[56] 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*, pages 3507–3516, 2019.

[57] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z Li, and Guoying Zhao. NAS-FAS: Static-dynamic central difference network search for face anti-spoofing. *arXiv:2011.02062*, 2020.

[58] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *CVPR*, pages 5295–5305, 2020.

[59] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face anti-spoofing database with diverse attacks. In *ICB*, pages 26–31. IEEE, 2012.