A Multi-Teacher Assisted Knowledge Distillation Approach for Enhanced Face Image Authentication

Tiancong Cheng
Northwestern Polytechnical University
Xi'an, Shaanxi, China
ctc1116@mail.nwpu.edu.cn

Ying Zhang*
Northwestern Polytechnical University
Xi'an, Shaanxi, China
izhangying@nwpu.edu.cn

Yifang Yin
Institute for Infocomm Research (I^{2}R), A*STAR
Singapore
yin_yifang@i2r.a-star.edu.sg

Roger Zimmermann
National University of Singapore
Singapore
rogerz@comp.nus.edu.sg

Zhiwen Yu*
Northwestern Polytechnical University
Xi'an, Shaanxi, China
zhiwenyu@nwpu.edu.cn

Bin Guo
Northwestern Polytechnical University
Xi'an, Shaanxi, China
guob@nwpu.edu.cn

ABSTRACT

Recent deep-learning-based face recognition systems have achieved significant success. However, most existing face recognition systems are vulnerable to spoofing attacks where a copy of the face image is used to deceive the authentication. A number of solutions are developed to overcome this problem by building a separate face anti-spoofing model, which however brings in additional storage and computation requirements. Since both recognition and face anti-spoofing tasks stem from the analysis of the same face image, this paper explores a unified approach to reduce the original dual-model redundancy. To this end, we introduce a compressed multi-task model to simultaneously perform both tasks in a lightweight manner, which has the potential to benefit lightweight IoT applications. Concretely, we regard the original two single-task deep models as teacher networks and propose a novel multi-teacher-assisted knowledge distillation method to guide our lightweight multi-task model to achieve satisfying performance on both tasks. Additionally, to reduce the large gap between the deep teachers and the light student, a comprehensive feature alignment is further integrated by distilling multi-layer features. Extensive experiments are carried out on two benchmark datasets, where we achieve the task accuracy of **93%** meanwhile reducing the model size by **97%** and reducing the inference time by **56%** compared to the original dual-model.

CCS CONCEPTS

• **Networks** → **Network performance analysis**.

*Corresponding author.

KEYWORDS

face recognition, face anti-spoofing, face authentication, model compression, knowledge distillation

ACM Reference Format:
Tiancong Cheng, Ying Zhang, Yifang Yin, Roger Zimmermann, Zhiwen Yu, and Bin Guo. 2023. A Multi-Teacher Assisted Knowledge Distillation Approach for Enhanced Face Image Authentication. In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3591106.3592280

1 INTRODUCTION

Face recognition is one of the most studied and successful biotechnologies in the past two decades thanks for its easy access, contact-free and high sensibility. We have witnessed that modern deep-learning-based face recognition models have achieved human-level performance[4, 24, 25]. These deep neural network-based models continuously raise the upper limit of face recognition tasks by stacking convolution modules and designing more suitable structures. However, most of these recognition systems are vulnerable to spoof attacks, which are carried out by malicious users to impersonate someone else's identity by using a copy of the photo or even videos[22]. To work against face spoofing attacks, most face authentication systems typically employ an additional module to analyze the spoofing status of the input faces before feeding it into the face recognition module [1, 21, 26]. Therefore, two deep models are typically deployed in a system to perform both face recognition (**FR**) and face anti-spoofing (**FA**) tasks which require additional computation and storage.

Since both FR and FA tasks are stemmed from the analysis of the same face image, this work unifies these two tasks by multi-task learning which has shown success in many other face-related tasks [20, 28, 30]. Additionally, considering that more and more **enhanced face authentication (FR plus FA)** vision systems are resource-constrained devices (e.g., campus access control, face payment equipment) and usually quick response is required [11, 35], it is therefore highly necessary to further lighten a usual multi-task model with a smaller network size and to improve the overall inference speed.

To achieve the above goals, this paper introduces a lightweight and unified approach to simultaneously perform both tasks by

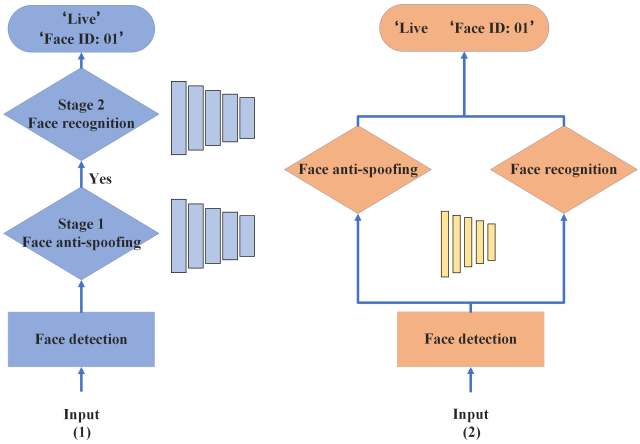

Figure 1: While prior studies use two separate and deep models to complete face recognition and face anti-spoofing in a sequential manner (left), we use a unified and compressed model to achieve both tasks simultaneously which is features of smaller size, less storage, easier training and faster inference speed (right).

using a compressed multi-task model. To achieve the satisfying performance of both tasks, instead of direct multi-task training, we take the original deep models as an explicit guide. Concretely, the original (two) single-task models are regarded as "teachers" while our model works are taken as a "student", and by distilling and transferring the knowledge from two teachers could our student model achieve satisfying performance for both tasks. Secondly, our student model is small in size while the original teacher models are significantly deep, if we adopt the output-level feature alignment between student and teacher(s), it will result in poor knowledge transfer. Therefore, we further propose a multi-layer feature alignment mechanism to comprehensively reduce such gaps. Our main contributions can be summarized as follows:

- To the best of our knowledge, it is the very first work to integrate face recognition and face anti-spoofing tasks in a single model by considering both task performance and model size. Compared to the traditional dual-model design, our unified model significantly reduces the storage requirement by **97%** and accelerates the inference time by **56%** with a slight performance drop of 2.5% on average.
- Our model is lightweight but accurate. This is achieved by our novel multi-layer multi-teacher assisted knowledge distillation (MMAKD) method which can comprehensively transfer the knowledge from original (two) deep single-task teacher networks to the lightweight multi-task student network. Compared to conventional direct multi-task training, our solution can boost the lightweight student's average performance of 5.7% and 4.9% in two datasets.
- Extensive experiments are carried out on two large-scale recent benchmarks and our model outperforms three state-of-the-art knowledge distillation methods by 4.8%, 5.4%, 4.9% and 1.7% on the average task accuracy, respectively.

2 RELATED WORK

Our work aims to achieve enhanced face authentication by leveraging knowledge distillation techniques. Therefore, this section reviews the mainstream and recent related works in both face authentication and knowledge distillation areas.

2.1 Face Authentication

Face authentication is proposed to overcome spoof attacks and improve the security and reliability of face recognition systems. The existing methods focus on independently solving the problems of face recognition and face anti-spoofing attacks for independent model building, training, and reasoning [1, 21, 26]. Through the analysis of these two-stage face authentication models, we summarize the relevant work into face recognition and face anti-spoofing from different stages of the model.

Face recognition methods. The mainstream research methods of face recognition based on machine learning can be divided into two categories: feature mapping method represented by FaceNet[25] and classification method represented by VggNet [24]. Some datasets [4, 31] introduce a new large-scale face dataset to improve the accuracy of face recognition by adding a diversity of poses and ages or adding a new environment. Khosla et al. [16] extend the self-supervised batch contrastive method to fully supervised scenes by clustering points of the same category into the embedding space, which can better utilize the label knowledge. Guo et al. [12] proposed to combine source/target domain conversion with a meta-optimization target to solve the face recognition problem in the unknown domain.

Face anti-spoofing methods. As CNN has proven to successfully outperform other learning paradigms in many computer vision tasks, deep-learning-based face anti-spoofing fields begin to be noticed by more and more researchers. Atoum et al. [2] propose the use of full convolutional neural networks to estimate the depth of an input face, the resulting depth map is then fed into the support vector machine to distinguish between real and deceptive faces. New convolutional neural networks are proposed in [6, 34], using the attention module to fuse the knowledge extracted from the complementary space RGB and MSR, or a two-side network to aggregate multi-level bilateral macro and micro information. Yu et al. [36] proposed a novel frame-level face anti-spoofing method based on central difference convolution, providing robust modeling capacity in describing detailed fine-grained information.

Enhanced Face Authentication. This work refers the enhanced face authentication as a combination of both recognition and face anti-spoofing, which is a topic that gradually received more attention. To this end, Albakri and Alghowinem [1] propose a two-stage model where the first step is enrolling the user on all devices and the second stage is to identify the spoofing status of the input images. Smith et al. [26] propose a scheme framework first extracts spoofing features and then realizes facial attribute feature extractions. Liu et al. [21] use a convolutional neural network to detect whether the image is live and then used FaceNet for face recognition.

Summary. From above, we observe that majority studies focus on treating the recognition and face anti-spoofing as separate models or modules. Though a few studies try to combine them, the integration is mostly at a system level and the two parts are placed

in a sequential manner. Different from these works, we design a multi-task single model as a one-stage face authentication and the two objectives is achieved in a parallel manner.

2.2 Knowledge Distillation

The knowledge distillation method has been widely used in the field of model compression recently, as it enables small student models to learn the ability of knowledge-rich teachers. Compressed models are necessary because most face verification tasks are deployed in resource-constrained scenarios. We next summarize some 1(teacher model)-to-1(student model) and n(teacher models)-to-1(student model) knowledge distillation methods.

In the 1-to-1 scenarios, HinTon et al. [14] proposed that the softmax distribution generated by the teacher's model is distilled into a soft target after a certain temperature treatment and it'll be learned by the student. Hou et al. [15] find that the front-layer feature visual attention captures a richer scene context, and it is proposed that the deep neural network learns the attention-based feature representation of the front layer. On the opposite, Chen et al. [7] propose a back-to-front fusion of various layers of characteristics of the student model based on the attention mechanism to make full use of the ability of advanced stages. He et al. [13] proposed the use of a non-parametric matrix singular value decomposition method to align the number of student and teacher feature channels in the field of intermediate feature layer distillation.

There are also a lot of works focused on improving the performance of compressed single-task students distilled from large single-task teachers [5, 9, 23, 27]. To solve the problem of mismatch between students and teachers, stop the teacher model update process early during training [9] or join the middle performance between large and small teacher assistant models [23] work well. Chen et al. [5] proposed that each student layer is automatically assigned the appropriate teacher model target layer through an attention mechanism to solve the right match between student and teacher knowledge. A patient distillation method of multiple intermediate layers is used in shallow single-task BERT models which is two times smaller than the teacher in [27].

In the n-to-1 scenarios, some works proposed to distill the multi-task student model with single-task pretrained teachers by the intermediate feature map outputs [19, 32]. Clark et al. [10] knowledge distillation where single-task models teach a multi-task model, the method is using teacher annealing of the logits to address that the student may be limited by the teacher's performance. Kundu et al. [17] proposed using cross-task distillation to reduce the discrepancy in the cross-task transfer domain, they first train a set of task-transfer networks and then carry out a multi-task adaptation framework. Walawalkar et al. [29] proposed to gradually compress the student model, and the compressed models at each stage are set into an ensemble. Then the smallest model is distilled by the online feature map outputs and the ensemble of logits.

Summary. While knowledge distillation has helped a compressed model to achieve a comparable good result as its full-sized teacher network, they mostly focus on improving the performance of the full-size model in a specific task. There are few works that discuss how knowledge distillation can be used for precision compensation under extreme compression or study multi-task compression distillation. Different from them, our work comes up with an n(teacher models)-to-1(multi-task student model) extremely compression distillation scenario. We propose a multi-layer shared-feature alignment mechanism to reduce the gaps between one tiny multi-task student and two large teachers simultaneously.

3 METHOD

In this section, we introduce the proposed lightweight multi-task model with a comprehensive feature alignment mechanism by integrating multi-layer features for face authentication, and the framework diagram is shown in figure 2. Technically, our work is composed of two parts: a lightweight multi-task student design module and a multi-teacher-assisted distillation module, a detailed description is as below.

3.1 Lightweight Multi-task Student Design

To integrate recognition and face anti-spoofing in a single model, we adopt multi-task learning as an attempt, which is inspired by its success on various face-related tasks [20, 28, 30]. Concretely, we use hard parameters sharing-based multi-task learning to reduce the overall parameters. However, since the shared blocks usually contain a large number of convolutional layers, the overall model size is still large. Therefore, we further compress the shared blocks by using the model pruning technique. Model compression methods can be divided into convolutional layer channel pruning, fully connected layer channel pruning, and shared block layer number pruning. Considering that network depth is an important factor to ensure the efficiency of visual tasks and the computational parameters of the model is mainly concentrated on convolutional layers. Thus the channel pruning method based on pruning filters for the convolutional layers method proposed by Li et al. [18] is used to get our lightweight student module. We obtain the weight of each channel in the convolutional layers through a powerful trained model, then the channels with smaller weights are pruned in a certain proportion. After fully measuring the final size of the pruned models and their accuracy, we determine the most appropriate pruning ratio of the shared modules based on the generalization ability of the dataset. A lightweight multi-task model with less than 1M parameters is obtained after this step.

$$\mathcal{L}_{mtl} = \alpha \mathcal{L}_{anti-spoofing} + \beta \mathcal{L}_{recognition} \tag{1}$$

The lightweight multi-task student model is optimized with Equation (1). $\mathcal{L}_{anti-spoofing}$ and $\mathcal{L}_{recognition}$ are the single-task loss. α and β are the multi-task trade-off hyperparameters that are set empirically. For convenience, we design the face anti-spoofing single-task branch with the exact same structure as the face recognition branch except for the final classification category (individuals number for the FR branch and number:*2* ('live ' and 'fake') for the FA branch). We use the cross-entropy loss for the FA single-task and details of the loss are as follows.

$$\mathcal{L}_{anti-spoofing} = \mathcal{L}^{ce}\left(\varphi\left(\cdot, \vartheta_s^{\tau}\right) \circ \phi\left(x, \theta_s\right), y_{\tau}\right) \tag{2}$$

The loss is calculated with the cross-entropy loss $\mathcal{L}^{ce}$ between the categorical probability output and the truth label $y_{\tau}, \tau = FA$.

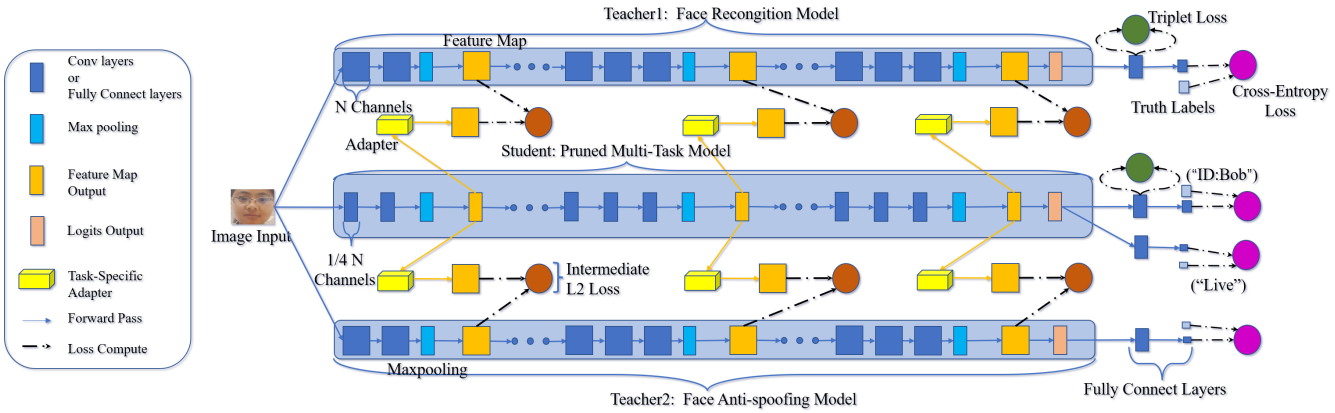

Figure 2: Framework diagram of our lightweight face authentication model. Teacher 1 (upper part) and Teacher 2 (lower part) are two deep models that are originally designed for face recognition and anti-spoof detection, respectively. The Student (middle part) is our unified multi-task model and could output a given face image's identity and spoofing status at the same time. For the lightweight purpose, we first establish the student as a compressed version of its teachers by channel-pruning. And to ensure a satisfying performance as teachers, we align the features among multiple shared layers to both teachers, which is realized by the task-specific adapters.

We treat all input images as x and its true labels y_τ for task τ, x is a triplet input containing three images (anchor a, positive p and negative n) in fact. $\phi\left(\cdot, \theta_s\right)$ encodes the input image into a high-dimensional encoding which represents the shared network's output. The task-specific predictor $\varphi\left(\cdot, \vartheta_s^\tau\right)$ takes in the shared encoding ϕ and outputs the logits for task τ, $i.e. \varphi\left(\cdot, \vartheta_s^\tau\right) \circ \phi\left(x, \theta_s\right)$, $\vartheta_s^\tau, \theta_s$ are the task-specific and share parameters of our model.

For face recognition single-task, we follow the face mapping 128-dimensional space proposed in FaceNet [25]. The FR loss function is a linear combination of triplet loss (formulation on the left) as well as cross-entropy loss $\mathcal{L}^{ce}$ which is used to accelerate convergence. $\mathcal{L}^{ce}$ is similar to equation (2) but $\tau = FR$. The loss function of our FR single-task is as follows ('re' represents 'recognition'):

$$\mathcal{L}_{re} = \sum_{a,p,n \epsilon \mathcal{D}} \left[max\left(\|a_{tri} - p_{tri}\|^2 - \|a_{tri} - n_{tri}\|^2 + \gamma, 0 \right) \right] + \mathcal{L}^{ce} \tag{3}$$

Three 128-dimensional feature mapping vectors a_{tri}, p_{tri} and n_{tri} are outputs of the feature mapping layer before the last classification fully-connect layer which takes in three images (anchor a, positive p and negative n) from dataset $\mathcal{D}$ simultaneously. The formulation before $\mathcal{L}^{ce}$ represents the triplet loss of the feature mapping module which allows the distance between positive examples in this euclidean space to be small enough and the negative examples' distance to be considerable. γ is a threshold set to 0.2.

3.2 Multi-Teacher Knowledge Distillation

From above, we establish a lightweight multi-task student network which is a trimmed version of its original full-size model. Smaller network size is a plus point from a resource-requirement aspect, however, our preliminary experiments show that direct multi-task training of this small multi-task network produces poor performance. Therefore, additional precision compensation is required and we explore the possibility to leverage knowledge distillation for this purpose. So the key issues include: 1) how to effectively distill from a deep teacher to a tiny student, and 2) how to distill knowledge from two teacher networks together.

Previous studies show that knowledge distillation usually fails when the gap between students and teachers is too large, [9, 23]. It is also found that the logits (logits are the soft probability output of the last linear fully connected layer) knowledge distillation method did not work on our lightweight multi-task student model as it doesn't have enough knowledge. On the other hand, the intermediate feature maps output from each max pooling layer often contains richer knowledge than the final fully connected layer's logits output for visual tasks [5]. Therefore, it would be helpful to integrate features from the intermediate layers as much as possible.

We know that the receptive field of a deep layer and a shallow layer in a neural network varies, and such difference is usually reflected on the different areas of interest of the layer-specific feature map. Figure 3 visualized such difference in our face authentication scenario where the feature maps are from the front, middle and last layer, respectively. The existence of such difference does not depend on the model type (i.e., either the single-task models such as the teacher network, or the multi-task model presents such difference), Therefore, instead of using a single intermediate layer's feature, we include the distillation from multiple layers. On the other hand, in order to ease the training, we do not distill from all the intermediate layers but choose the representative layers only. In our work, the number of layers that need to be distilled is set to $I = 3$ under the simplicity of tuning hyper-parameters and the choice of layers was detailed discussed in the Experiment Section.

Before the distillation, we first train the two large teacher models (one for the recognition task and the other for the liveness detection task) with pretrained parameters on CASIA-WebFace [33] by optimizing Equation(2) and Equation(3), separately. This part is conducted in an offline manner. During the distillation, since

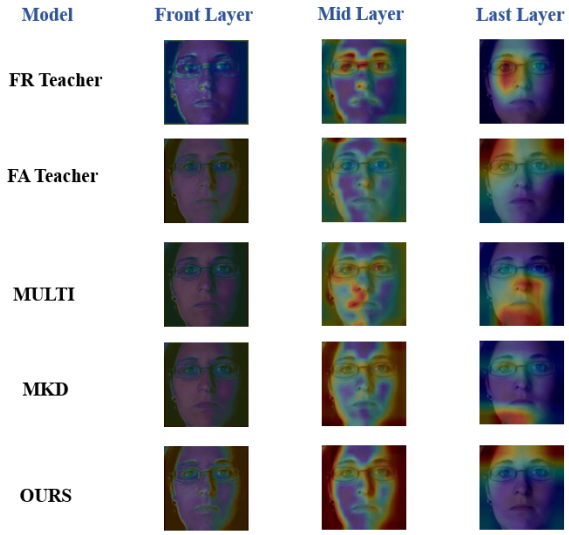

Figure 3: Visualization of the front-mid-last multi layers' feature map outputs using CAM [38] from multiple baseline models. From the first three lines of the results are easy to see, the multi-task model's output includes the preference of the FR task attention map for the eyes and the FA task for global information simultaneously. From the last two lines, it can be seen that the feature maps output from our proposed MMAKD model incorporates more refined attention to facial area than the MKD model.

our compressed student model's feature encoders have different shapes of the output from the teachers, we propose an adapter for an effective feature alignment. Since there are two teacher networks, we name the adapter from the student to each teacher as a task-specific adapter. In total, there are six task-specific adapters A_τ^i ($\tau \epsilon T$=[FR, FA] , $i \epsilon I$) to project student feature encoder's outputs $\phi\left(x, \theta_s^i\right)$ into each task-specific one: $\mathbb{R}^{C_{si} \times H_{si} \times W_{si}} \Rightarrow \mathbb{R}^{C_{ti} \times H_{ti} \times W_{ti}}$ where C, H, W are the channels' number, height, and width of the feature maps. Each adapter uses a linear layer consisting of a $3 \times 3 \times C_{in} \times C_{out}$ convolution. The specific knowledge distillation loss is Equation(4):

$$\mathcal{L}_{kd} = \sum_{\tau}^{T=2} \sum_{i}^{I=3} \lambda_\tau^i \mathcal{L}^d \left(A_\tau^i \left(\phi \left(x, \theta_s^i \right) \right), \phi \left(x, \delta_\tau^i \right) \right) \tag{4}$$

where $\phi\left(x, \delta_\tau^i\right)$ is to represent the i_{th} feature map output (output of each max pooling layer) from τ_{th} teacher, δ_τ^i represents the i_{th} shared layer parameters of teacher τ. $\mathcal{L}^d$ is the Euclidean distance function (equation 5) to calculate the distance between the L2 normalized feature maps, λ_τ^i are the hyperparameters that need to be determined empirically. The distance is calculated as in Equation5:

$$\mathcal{L}^d \left(f_1, f_2 \right) = \left\| \frac{f_1}{\|f_1\|_2} - \frac{f_2}{\|f_2\|_2} \right\|_2^2 \tag{5}$$

where $f^i, f^j \in \mathbb{R}^{C \times H \times W}$ and C, H, W indicates the channels' number, height and width of the feature maps by the feature encoders $A_\tau^i\left(\phi\left(x, \theta_s^i\right)\right)$ and $\phi\left(x, \delta_\tau^i\right)$. The overall loss of the model is a linear combination of Equation(1) and Equation(4).

$$\mathcal{L} = \mathcal{L}_{mtl} + \mathcal{L}_{kd} \tag{6}$$

Last but not least, the distillation performance is related to the model compression level. From our experiments, it seems that a student models with less than 1M parameters is not able to learn very useful knowledge by a direct easy training (only using the task's ground truth labels). However, the small models have sufficient capacity to imitate from the teacher model. We will analyze the relationship between distillation strength and model compression ratio in Experiment Section.

4 EXPERIMENTS

4.1 Dataset

To demonstrate the effectiveness of our MMAKD model, we performed experiments on two common benchmark datasets which contain both people identities and the face spoofing labels, including Oulu-Npu and Replay-Attack.

Oulu-Npu [3] consists of 4950 real access and attack video attempts to 55 clients, recorded by six mobile devices, and four different kinds of printing and video replaying face attacks. All videos involve three different lighting conditions and three different backgrounds which have a higher generalized ability and are harder to study. We extract images from video frames and then crop them using MTCNN [37] to ensure the validity of the knowledge learned by the model. The number of final training datasets and test datasets was 3058 containing 35 face identities and 1728 containing 20 face identities respectively. The ratio of real face pictures to fake faces is approximately 1:1.

Replay-Attack [8] consists of 1300 video clips of photo and video attack attempts to 50 clients, the spoofing attacks are divided into printing attacks and displaying attacks by iPad and iPhone. During our experiments, we found that the fake face data in this dataset was not generalized enough (there were only four types of face attacks) and this situation leads to a very high degree of accuracy in the FA single-task branch. Our training dataset contains 2400 images from 30 face identities, on the contrary, the test dataset contains 1600 images from 20 face identities. The ratio of real face pictures to fake faces is approximately 1:4.

4.2 Baselines

The baselines we used are as follows:

STL: STL refers to the single-task learning models for either FR or FA based on the pre-trained VGG-16 networks. The two STL models are regarded as large teachers.

1/8 MTL: 1/8 MTL refers to our lightweight multi-task model based on the hard parameters sharing architecture, which needs to be taught. The fraction '1/8' means 87.5% channels of the multi-task model's shared block are pruned.

LKD [14]: Students learn the teacher's soft logits at a certain temperature and the truth label knowledge.

BAM [10]: Students learn the teacher's soft logits and the truth label knowledge with an annealing weight.

MKD [19]: Student learns from the teacher's feature maps by the euclidean distance between the two and the truth label knowledge.

Table 1: Testing accuracy on Oulu-Npu / Replay-Attack Datasets.

Type	Methods			Face Recognition	Face Anti-spoofing		Average	Inference
	Model	*Distill*	*Train Params*	*Accuracy*	*Accuracy*	*HTER*	*Accuracy*	*Time*
STL	FR	-	29.55M	0.952 / 0.941	-	-	0.952 / 0.966	4.8 / 5.1ms
	FA	-		-	0.952 / 0.990	0.061 / 0.007		
1/8 MTL	MTL	-	0.25M	0.865 / 0.819	0.883 / 0.955	0.135 / 0.045	0.874 / 0.887	2.0 / 2.1ms
	LKD [14]	Yes	0.25M	0.888 / 0.812	0.878 / 0.927	0.211 / 0.070	0.883 / 0.870	
	BAM [10]	Yes	0.25M	0.885 / 0.814	0.869 / 0.918	0.216 / 0.080	0.877 / 0.866	
	KR [7]	Yes	1.99M	0.891 / 0.835	0.873 / 0.908	0.180 / 0.099	0.882 / 0.872	
	CKD [5]	Yes	25.48M	0.916 / 0.867	0.930 / 0.984	0.090 / 0.017	0.923 / 0.926	
	MKD [19]	Yes	0.85M	0.919 / 0.856	0.909 / **0.992**	0.115 / **0.008**	0.914 / 0.924	
	Ours	**Yes**	0.99M	**0.928** / **0.882**	**0.933** / *0.990*	**0.087** / *0.010*	**0.931** / **0.936**	

Different from MKD, the student model in our work has a huge chasm with teachers' capability since it is almost 14 times smaller than the teacher model and their teachers are the same as their student network instead.

KR [7]: KR is an attention based feature map distillation method. The feature maps of students are fused based on the attention mechanism from the back to the front.

CKD [5]: CKD is applied in compression distillation scenarios. It utilizes attention mechanisms to determine distillation weights between one student layer and different teacher layers. The method imposes the same attention projection on teachers' features as student's features, and the amount of parameters is determined by the teachers' features' dimension.

4.3 Metrics

We choose widely used metrics to conduct experiments: triplet accuracy for the FR task, Top-1 accuracy and HTER for the FA task:

Triplet accuracy: during the testing of the model, a suitable threshold is determined by k-fold cross-validation, which is used to determine how many triplet samples can meet the goal of the model's 128-dimensional feature space mapping function: anchor's distance from positive cases are small enough and anchor's distance from negative sample samples are large enough.

HTER is the half total error rate, which is half of the sum of the proportions of real and fake faces that are judged wrong.

4.4 Settings

Our multi-task learning model (figure 2) is trimmed 87.5% of the convolution channels from the shared blocks, and the weights α and β of the two single tasks are all set to 0.5. Data in the form of triples were used and the batch size is set to 66. We set the hyperparameter λ_τ^i to [1,1,8] for each task based on the experience. We use SGD for optimizing the networks and adaptors and the momentum is set to 0.9. The learning rate of all optimizers (including task-specific adaptors) is set to 0.01. We train single-task models for 30 epochs and all multi-task models for 60 epochs.

4.5 Results and Analysis

Results on Oulu-Npu. Table 1 shows that the MMAKD model which distills multi-layer feature maps achieves the best performance, the accuracy of both FR and FA single-task branches is close to the best teachers' results which take pre-trained parameters and no channel trimming. The final experimental results prove that the simultaneous distillation of the deep and shallow features in a comprehensive way can guarantee impressive improvement of the distillation from large teachers to the small student (we call it **large-to-tiny distillation**). Compared with the original dual model, the inference time is reduced by 56% and the average accuracy has a slight performance drop of 2.1% on average conversely. Compared to conventional direct multi-task training, our solution can boost task performance by 6.3% and 5.0%.

The result of LKD and BAM proves that the logits distillation has played a weak role in terms of performance improvement in large-to-tiny scenarios. Attention-based method KR did not work well because there were redundant parameters that needed to be optimized, which leads to the transfer of invalid knowledge during the distillation process. The CKD method achieves similar performance to ours as it distills the same layers as we do, but the training parameters and computation far exceeds ours. The dynamic distillation weights derived from the attention mechanism do not play a significant role in our large-to-tiny scenario. The comparison with MKD method also reflects that knowledge from different stages of the teachers has different effects on the accuracy balance of multi-task models. Therefore, the balance of multi-task models can also be achieved by distilling knowledge from different layers.

Results on Replay-Attack. The experimental setup on this dataset is the same as for Oulu-Npu. We train with the full training set and report the results of two tasks on the test set in Table 1. The results on this dataset are similar to the above, and the MMAKD model we proposed accelerates more significantly in inference time. Since the generalization ability of this dataset on FA single task is high enough, the performance improvement of the MMAKD model is not obvious on FA single task, but it is still 1% higher than the sub-optimal MKD model in terms of average accuracy. Compared to conventional direct multi-task training, our solution can boost task performance by 6.3% and 3.5%. The inference time is reduced by 59% and the average accuracy has a slight performance drop of 3% on average compared with the original dual model.

Analysis on the inference time. After experimentation, we find that the inference time of the model is not linear with the number of parameters. For memory-intensive operators (Relu activation function), the inference time is linearly related to the access stock,

while for computationally intensive operators (Conv and FC functions), the inference time is linearly related to the computation quantity. Our compression model greatly reduces the number of parameters and computations, thus the slight speed improvement is worthwhile. We will further explore modifying memory-intensive operators to accelerate the inference speed in future experiments.

4.6 Feature Map Visualization

To further understand the difference among the features from each layer of the neural networks during the knowledge distillation process, and prove the effectiveness of our proposed comprehensive feature alignment method. We use CAM [38] to the feature map output of each layer of the neural network. We visualized the feature map output after each max pooling layer from FR and FA single-task teacher models' shared blocks, the others are from the proposed lightweight multi-task model and the best two multi-task distillation models. The attention maps are shown in figure 3, and the average accuracy of the model corresponding to the last three rows gradually increases. It is easy to see that the multi-task model's output in the general sense is the organic combination of two single-task model feature maps, which includes the preference of the FR task attention map for the eyes and the FA task for global information simultaneously. Higher accuracy means more similar the multi-task model feature layer output is to the perfect fusion of the corresponding layer outputs of two teacher models. The MTL model has a false focus on the middle region, and the MKD method does not pay enough attention to facial features. Our proposed model reflects the model's more refined attention to images and achieves the best performance.

5 ABLATION STUDY AND DISCUSSION

In order to better understand the application of the knowledge distillation method in the scenario of a very small student model distilled from the large teacher (we call it **large-to-tiny distillation**), we conduct some ablation experiments on the Oulu-Npu dataset. First, we report the performance of the multi-layer multi-teacher assisted knowledge distillation (MMAKD) method under different compression ratio models and different distillation layer numbers. The results show that our method is very useful in the scenario of a small student model distilled from large teachers, and the effect is more significant when the compression ratio of the student model is larger. Then, through a large number of experiments, we have demonstrated that students with a small number of parameters should devote all their abilities to learning teachers in the distillation process and that the weight of distillation loss (λ) should be much higher than that of multi-task loss (α and β).

Analysis of the number of distillation layers. To demonstrate the effectiveness of our method, we analyze the influence of the number of distillation layers on student models with different compression ratios in the large-to-tiny distillation scenario. In this experiment, we set the distillation weights λ_τ^i for each layer to equal simplicity (*e.g.* [1,1,1,1,1]) in one experiment, which is different from the best hyperparameters set in Section 4. The MTL model with VGG-16 as the backbone has a total of five max pooling operations, corresponding to five feature map outputs (layer [0,4]). In the second column of Table 2, 'front' stands for layer 0 output, 'mid' represents layer 2 output, 'last' is layer 4 output, and 'all layers' means distill feature map outputs from layer 0 to layer 4. We compress the multi-task model using the channel pruning algorithm mentioned in Section 3. '1/4 MTL' represents 75% channels in shared blocks of the MTL model are trimmed. Different fractions before the 'MTL' symbol stand for different compression ratios.

Table 2: Ablation study on the number of distilling layers. Here, 'FR' means face recognition accuracy, 'FA' means face anti-spoofing accuracy, and 'AVG' means average accuracy. '✗' means no layers to be distilled. The fraction before 'MTL' represents the proportion remaining after the convolutional channels in the shared module has been pruned, and its parameter amount is in parentheses.

Type	Distill Layers	FR	FA	AVG
MTL (14.85M)	✗	0.926	0.905	0.916
	last	0.946	0.944	0.945
	mid+last	0.934	0.952	0.943
	front+mid+last	0.939	0.957	**0.948**
	all layers	0.923	0.954	0.938
1/2 MTL (3.75M)	✗	0.909	0.919	0.914
	last	0.937	0.949	0.943
	mid+last	0.925	0.961	0.943
	front+mid+last	0.927	0.963	0.945
	all layers	0.936	0.957	**0.946**
1/4 MTL (0.96M)	✗	0.911	0.897	0.904
	last	0.917	0.937	0.927
	mid+last	0.912	0.947	0.929
	front+mid+last	0.918	0.951	**0.934**
	all layers	0.917	0.949	0.933
1/8 MTL (0.25M)	✗	0.865	0.883	0.874
	last	0.919	0.909	0.914
	mid+last	0.919	0.921	0.920
	front+mid+last	0.928	0.933	**0.931**
	all layers	0.910	0.944	0.927

From the results shown in Table 2, it's easy to see that the MMAKD method proposed by us works well in various distillation types, especially the large-to-tiny distillation type. In the '1/8 MTL' type, the average accuracy of distilling three layers is 1.7% higher than that of distilling single layers and 5.7% higher than that of multi-task models without distillation.

Future analysis of distillation weights. For further exploring efficient solutions of large-to-tiny distillation scenarios, we perform different degrees of knowledge distillation on models with different compression ratios. The compression ratios are taken at 0%, 25%, 50%, 75%, and 87.5% (full, 3/4, 1/2, 1/4, and 1/8 MTL), which are applied in the shared blocks of our full-size MTL model. We use the single-layer distillation method (MKD [19]) to simplify the parameters tuning process. Distillation weights of the single-layer distillation MTL model are reflected in the weights λ_1, λ_2 that control the strength of loss $\mathcal{L}_{kd}$. We choose six kinds of distillation

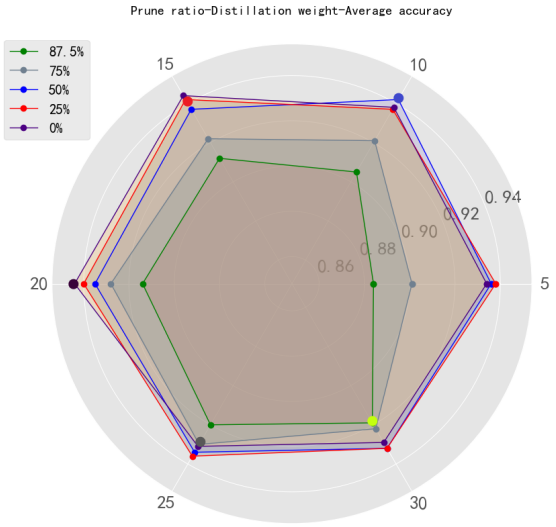

Figure 4: Radar plot related to the degree of compression of the model as well as the strength of distillation. Different colors represent different compression ratios, the outer ring numbers represent different distillation weights, and the inner ring numbers represent the multi-task average accuracy.

weights {[2.5,2.5], [5,5], [7.5,7.5],[10,10],[12.5,12.5],[15,15]}, reflecting in the numbers {5,10,15,20,25,30} on the periphery of the radar chart (figure 4). This analysis result is shown in figure 4.

The vertices of the colored polygons represent the average accuracy of the multi-task model under a certain compression ratio and distillation strength. We highlight the best distillation results for each compression ratio. As can be seen from the figure, in scenarios where the parameters of the student model are large (compression ratio is low), the distillation intensity should be as large as possible to get better performance. For example, 20 ([10,10]) for full-size MTL can reach 0.944 average accuracy, and 15 ([7.5,7.5]) for 25%-pruned-size MTL can reach 0.942 average accuracy. For medium-sized student models (50% pruned of shared blocks' channels), the parameters of it are not so large, the distillation weights should be small too: 10 ([5,5]) to get the best average accuracy of 0.942. However, in large-to-tiny scenarios, the distillation strength should be extremely large. The distillation weight should be 25 ([12.5,12.5]) for 75%-pruned-size MTL to reach 0.930 average accuracy, and 30 ([15,15]) for 87.5%-pruned-size MTL to get an average accuracy of 0.920, which make distillation loss far exceeds the multi-task loss.

From table 2 and figure 4, we have the following observations:

- The multi-layer multi-teacher assisted knowledge distillation (MMAKD) method works well in large-to-tiny distillation scenarios, especially when the parameters of teacher models are 100 times larger than the lightweight multi-task student model (full-size teachers and 87.5%-pruned student).
- When the parameters of the student model are not much different from that of the teacher, the student has the extra ability to learn knowledge from truth labels that are not available in the teacher while learning the knowledge of distillation. However, the tiny student model should pay full attention to learning the knowledge of big teachers without learning the truth labels when the number of its parameters is extremely small. In this way, order ensures that the only abilities that primary school students have can be used to learn the richest knowledge from teachers.
- By properly adjusting the distillation weights or using multi-layer distillation, the middle student model can achieve an infinitely close ability to the large student model. For example, 1/2 MTL in table 2 achieves an average accuracy of 0.946, and 25% pruned MTL in figure 4 achieves an average accuracy of 0.942. The two only lost an accuracy of 0.002 compared to the best full-size model.

6 CONCLUSION

This work proposes a knowledge distillation-based lightweight face authentication model, solving face recognition and face anti-spoofing tasks at the same time. The use of a multi-task learning method and channel pruning algorithm reduces the inherent redundancy of the original dual model, accelerates the running speed of the model more than twice times and reduces the number of model parameters by more than 28M. We perform comparison experiments on public datasets and our comprehensive feature alignment method by simultaneously distilling knowledge from multi-layer feature maps. The results show that our knowledge distillation method has a higher average accuracy (5.7% and 4.9%) than the multi-task baseline without distillation and it achieves the best performance compared to state-of-the-art knowledge distillation methods at large teachers distill tiny multi-task student scenarios.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (No. 61960206008, 62272390, 62025205, 62072375, 62032020), and Singapore Ministry of Education Academic Research Fund Tier 2 under MOE's official grant number T2EP20221-0023.

REFERENCES

[1] Ghazel Albakri and Sharifa Alghowinem. 2019. The effectiveness of depth data in liveness face authentication using 3D sensor cameras. *Sensors* 19, 8 (2019), 1928.
[2] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. 2017. Face anti-spoofing using patch and depth-based CNNs. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 319–328.
[3] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. 2017. 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)*. IEEE, 612–618.
[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 67–74.
[5] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. 2021. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7028–7036.
[6] Haonan Chen, Guosheng Hu, Zhen Lei, Yaowu Chen, Neil M Robertson, and Stan Z Li. 2019. Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Transactions on Information Forensics and Security* 15 (2019), 578–593.
[7] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. 2021. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5008–5017.
[8] Ivana Chingovska, André Anjos, and Sébastien Marcel. 2012. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*. IEEE, 1–7.

[9] Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4794–4802.

[10] Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D Manning, and Quoc Le. 2019. BAM! Born-Again Multi-Task Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5931–5937.

[11] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. 2019. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[12] Jianzhu Guo, Xiangyu Zhu, Chenxu Zhao, Dong Cao, Zhen Lei, and Stan Z. Li. 2020. Learning Meta Face Recognition in Unseen Domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Ruifei He, Shuyang Sun, Jihan Yang, Song Bai, and Xiaojuan Qi. 2022. Knowledge distillation as efficient pre-training: Faster convergence, higher data-efficiency, and better transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9161–9171.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *stat* 1050 (2015), 9.

[15] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. 2019. Learning lightweight lane detection cnns by self attention distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1013–1021.

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18661–18673. https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf

[17] Jogendra Nath Kundu, Nishank Lakkakula, and R Venkatesh Babu. 2019. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1436–1445.

[18] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. [n. d.]. Pruning Filters for Efficient ConvNets. In *International Conference on Learning Representations*.

[19] Wei-Hong Li and Hakan Bilen. 2020. Knowledge distillation for multi-task learning. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 163–176.

[20] Shikun Liu, Edward Johns, and Andrew J Davison. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1871–1880.

[21] Shuhua Liu, Yu Song, Mengyu Zhang, Jianwei Zhao, Shihao Yang, and Kun Hou. 2019. An identity authentication method combining liveness detection and face recognition. *Sensors* 19, 21 (2019), 4733.

[22] Zuheng Ming, Muriel Visani, Muhammad Muzzamil Luqman, and Jean-Christophe Burie. 2020. A survey on anti-spoofing methods for facial recognition with rgb cameras of generic consumer devices. *Journal of Imaging* 6, 12 (2020), 139.

[23] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 5191–5198.

[24] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. (2015).

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[26] Max Smith-Creasey, Fatema A Albalooshi, and Muttukrishnan Rajarajan. 2018. Continuous face authentication scheme for mobile devices with tracking and liveness detection. *Microprocessors and Microsystems* 63 (2018), 147–157.

[27] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient Knowledge Distillation for BERT Model Compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4323–4332.

[28] Manh Tu Vu, Marie Beurton-Aimar, and Serge Marchand. 2021. Multitask multi-database emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3637–3644.

[29] Devesh Walawalkar, Zhiqiang Shen, and Marios Savvides. 2020. Online ensemble model compression using knowledge distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*. Springer, 18–35.

[30] Lingfeng Wang, Jin Qi, Jian Cheng, and Kenji Suzuki. 2022. Action unit detection by exploiting spatial-temporal and label-wise attention with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2470–2475.

[31] Zhongyuan Wang, Baojin Huang, Guangcheng Wang, Peng Yi, and Kui Jiang. 2023. Masked face recognition dataset and application. *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2023).

[32] Jingwen Ye, Xinchao Wang, Yixin Ji, Kairi Ou, and Mingli Song. 2019. Amalgamating filtered knowledge: learning task-customized student from multi-task teachers. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 4128–4134.

[33] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. [n. d.]. Learning Face Representation from Scratch. ([n. d.]).

[34] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. 2020. Face anti-spoofing with human material perception. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 557–575.

[35] Zitong Yu, Yunxiao Qin, Xiaqing Xu, Chenxu Zhao, Zezheng Wang, Zhen Lei, and Guoying Zhao. 2020. Auto-fas: Searching lightweight networks for face anti-spoofing. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 996–1000.

[36] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. 2020. Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5295–5305.

[37] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.