PipeNet: Selective Modal Pipeline of Fusion Network for Multi-Modal Face Anti-Spoofing

Qing Yang[1], Xia Zhu[2], Jong-Kae Fwu[2], Yun Ye[1], Ganmei You[1], and Yuan Zhu[1]

[1]Intel, China
[1]{qing.y.yang, yun.ye, ganmei.you, yuan.y.zhu}@intel.com
[2]Intel, USA
[2]{xia.zhu, jong-kae.fwu}@intel.com

Abstract

Face anti-spoofing has become an increasingly important and critical security feature for authentication systems, due to rampant and easily launchable presentation attacks. Addressing the shortage of multi-modal face dataset, CA-SIA recently released the largest up-to-date CASIA-SURF Cross-ethnicity Face Anti-spoofing(CeFA) dataset, covering 3 ethnicities, 3 modalities, 1607 subjects, and 2D plus 3D attack types in four protocols, and focusing on the challenge of improving the generalization capability of face anti-spoofing in cross-ethnicity and multi-modal continuous data. In this paper, we propose a novel pipeline-based multi-stream CNN architecture called PipeNet for multi-modal face anti-spoofing. Unlike previous works, Selective Modal Pipeline (SMP) is designed to enable a customized pipeline for each data modality to take full advantage of multi-modal data. Limited Frame Vote (LFV) is designed to ensure stable and accurate prediction for video classification. The proposed method wins the third place in the final ranking of Chalearn Multi-modal Cross-ethnicity Face Anti-spoofing Recognition Challenge@CVPR2020. Our final submission achieves the Average Classification Error Rate (ACER) of 2.21 ± 1.26 on the test set.

1. Introduction

Face recognition technology has been widely used in a variety of applications in our daily life, such as mobile face payment, entrance authentication and office check-in machines, etc. Unfortunately, human face's easy accessibility brings not only convenience but also presentation attacks. These facial recognition systems are vulnerable to various types of presentation attacks, including printed photograph, digital video replay, 3D print mask, and silica gel face *et*

al. Therefore, how to detect these various means of presentation attacks has become an increasingly critical and challenging task in all face recognition and authentication systems.

Various face anti-spoofing methods have been proposed over the past decade. Traditional methods use the combination of handcrafted features [30, 5, 14, 23, 37, 26, 13, 21, 7, 4] such as Sparse Low Rank Bilinear (SLRB), Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Difference of Gaussian (DoG), Scale Invariant Feature Transform (SIFT), Speeded-Up Robust Feature (SURF) and traditional classifiers including Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) to solve the classification problem of real and fake face. However, due to poor generalization capability, it is difficult to apply these methods in changing illumination environments.

Recently, deep convolutional neural networks (CNNs) have been widely used for various face related tasks such as face detection, recognition, identification and landmark detection *et al*. Researchers have also introduced CNNs into face anti-spoofing and liveness detection area. It is proved that CNNs [8, 16, 27, 35] can effectively extract richer deep facial image semantic features than traditional methods for binary classification of face anti-spoofing tasks. However, the single modal data input and supervised-learning mechanism makes it difficult to perform well in cross-dataset tests.

RGB data can provide texture information, but it is insensitive to surface curvature and reflectivity of materials. To improve the accuracy of face anti-spoofing systems (FAS), more modalities data such as depth [2] and InfraRed (IR) are used as algorithm data input. Depth camera like Intel RealSense D400 series[1] can provide distance/depth information to make it easier to distinguish real face surfaces

_____
[1]https://www.intelrealsense.com/

Figure 1. Low-cost high-precision 3D face masks.

from face shapes in photograph and video replays on electronic displays. IR camera is sensitive to material reflectivity of different presentation attack instruments (PAI). Especially, fusion methods [25, 28, 40] become main stream with the large-scale multi-modal dataset CASIA-SURF [41] publicly released.

Unfortunately, with the rapid development of presentation attack methods, face anti-spoofing systems have to deal with more challenging situations. As shown in Fig 1, it becomes easier and cheaper to produce 3D printed masks, while it's extremely difficult to distinguish them from real faces in both RGB and depth data modalities. In addition, silica gel face is more lifelike and it looks like real face in all 3 modalities (RGB, depth and IR). For most previous approaches, it's impossible to distinguish between these two presentation attacks and live face.

To address this issue, we propose a novel pipeline-based fusion framework which can effectively and respectively extract richer information from face data in different modalities.

In summary, the main contributions of this paper are summarized as follows:

(1) We present an effective pipeline-based framework PipeNet for multi-modal face anti-spoofing task;

(2) In contrast to previous unified designs, we propose a Selective Modal Pipeline (SMP) design for differentiating feature extraction among different modalities data;

(3) We propose a novel method - Limited Frame Vote (LFV) to get stable and reliable prediction for video classification.

The rest of the paper is organized as follows: Section 2 briefly reviews related works in multi-modal face anti-spoofing; Section 3 introduces CASIA-SURF CeFA dataset and baseline methods; Section 4 proposes our method PipeNet with SMP and LFV modules; Section 5 shows competition basic rules; Section 6 shows the experiments and results analysis; Finally, we summarize our work and future research direction in Section 7.

2. Related Work

Traditional Methods: The facial motions in a sequence of frames are firstly utilized as cues in face anti-spoofing task.

For example, eyes blinking is a reliable evidence that the face is real [29, 24].

In contrast to requesting user's cooperation, some existing approaches treat face anti-spoofing task as a binary classification problem which can proceed with single frame. Various hand-crafted features have been explored and adopted in previous works, such as SLRB, LBP, HOG, DOG, SIFT, and SURF. Followed by traditional classifiers such as SVM, LDA and Random Forest.

CNN-based Methods: After AlexNet's [15] great success, the strong representation capability of CNNs has been exploited in face anti-spoofing research [8, 16, 27, 36]. CNN methods [35] attempt to learn deep feature representations from face image data for binary classification. Two-stream CNNs [2] are proposed to overcome the weakness of cross-dataset test performance based on both patch and depth streams. Li [16] proposed a method to link partial features and Principle Component Analysis (PCA), instead of fully-connected layer, to reduce the dimensionality of features in order to avoid the over-fitting problem.

Temporal-based Methods: Temporal information is also used to enhance CNN's representation capability. Long Short Term Memory (LSTM) [38, 34] can recurrently learn features to obtain context information, but the heavy computation overhead makes it difficult to deploy such method, especially on mobile devices. Remote photoplethysmography (rPPG) [3, 22, 19] measures a pulse signal of heart rate which only exists in live face, and can be used to 3D masks attack. However, weak robustness makes rPPG vulnerable to illumination changes of environment. More importantly, both methods take a long time for one testing shot, which is unpractical for deployment.

Fusion-based Methods: Multi-Modal methods can take advantage of complementary information among different modalities data and provide more robust performance on face anti-spoofing task. However, there used to be only small face anti-spoofing dataset with limited subjects and samples available, which make it easy to fall into overfitting problem during CNN training.

In 2019, Zhang [41] released CASIA-SURF, a large-scale multi-modal dataset for face anti-spoofing, which consists 1000 subjects with 21,000 videos and each sample with 3 modalites data(*i.e.*, RGB, depth and IR).

With CASIA-SURF dataset, they hold the first round of Face Anti-spoofing Attack Detection Challenge@CVPR2019 [1]. The top-3 solutions were as follows: VisionLab [25] was the champion with ACER score of 0.0810%. They proposed a fusion network with multi-level feature aggregation module which can fully utilize the feature fusion from different modalities at both coarse and fine levels. FaceBagNet [28] won the second place with ACER score of 0.0985%. They proposed a patch-based multi-stream fusion CNN architecture based on Bag-

Subset	Ethnicities			Subjects	Modalities	PAIs	# real videos	# fake videos	# all videos
	4@1	4@2	4@3						
Train	A	C	E	1-200	R&D&I	Replay	600/600/600	600/600/600	1200/1200/1200
Valid	A	C	E	201-300	R&D&I	Replay	300/300/300	300/300/300	600/600/600
Test	C&E	A&E	A&C	301-500	R&D&I	Print	1200/1200/1200	5400/5400/5400	6600/6600/6600

Table 1. CASIA SURF CeFA dataset Protocol 4.

of-local-features. The patch-level images contribute to extract spoof-specific discriminative information and Model Feature Erasing module randomly erases one modal to prevent overfitting. FeatherNets [40] was the third winner with ACER score of 0.1292%. They proposed a light-weighted network architecture with modified Global Average Pooling(GAP) named streaming module. Their fusion procedure is based on *ensemble + cascade* structure to make best use of each modal data.
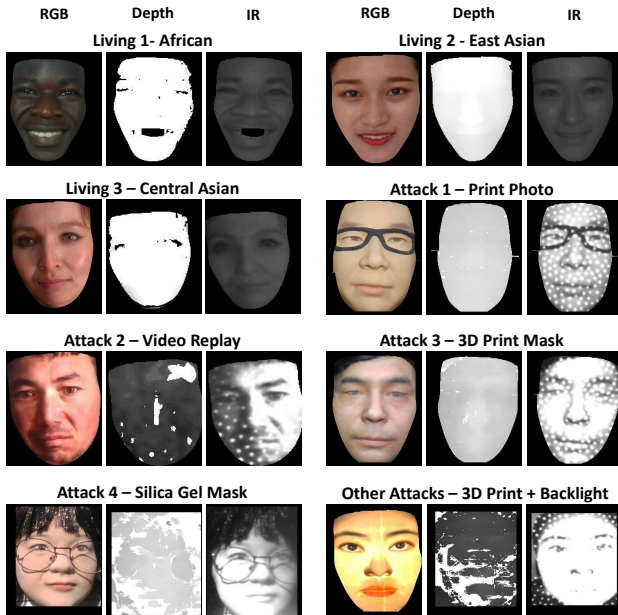

Figure 2. Examples of living and spoofing faces among different ethnicities and different attack types in three modalities (RGB, Depth, IR) from CASIA SURF-CeFA dataset [18].

3. Dataset And Baseline

3.1. CASIA-SURF CeFA

CASIA-SURF dataset only contains faces in one ethnicity. In order to provide a benchmark platform for improving the generalization capability in *cross-ethnicity* anti-spoofing and the utilization of multi-modal continuous data, CASIA further released the largest up-to-date *cross-ethnicity* face anti-spoofing dataset, CASIA-SURF CeFA [18] in 2020. CASIA-SURF CeFA covers 3 ethnicities, 3 modal-

ities, $1,607$ subjects, and 2D plus 3D attack types. It is the first public dataset designed for exploring the impact of cross-ethnicity in the study of face anti-spoofing. As shown in Figure 2, samples from different ethnicities and several presentation attack types are included in the CASIA-SURF CeFA dataset. Specifically, 4 protocols are introduced to measure the impact under various evaluation conditions: cross-ethnicity (Protocol 1), (2) cross-PAI (Protocol 2), (3) cross-modality (Protocol 3) and (4) cross-ethnicity&PAI (Protocol 4 as shown in Table 1). This paper is focus on most difficult one, Protocol 4.

3.2. Baseline Methods

SD-Net. CASIA provides a baseline for single-modal face anti-spoofing task via a ResNet-18-based [10] network named SD-Net [18]. It includes 3 branches: static, dynamic and static-dynamic branches to learn hybrid features from static and dynamic images, respectively. For static and dynamic branches, each of them consists of 5 res-blocks and 1 GAP layer. A detailed description is provided in [18] for how to generate dynamic images. In short, a dynamic image is computed online with rank pooling based on K consecutive frames. The reason for selection of dynamic images for rank pooling in SD-Net is that dynamic images have proved to have advantage over conventional optical flow [31, 9].

PSMM-Net. To make full use of multi-modal image data to alleviate the ethnic and attack bias, CASIA proposes a novel multi-modal fusion network, namely PSMM-Net [18], which includes two main parts: a) the modality-specific network, contains three SD-Nets to respectively learn deep features from three modalities such as RGB, Depth, IR; b) a shared branch for all modalities, for the purpose of learning the complementary features among different modalities. They also design information exchange and interaction among SD-Nets and the shared branch, with the purpose of capturing correlations and complementary semantic information among different modalities.Two main kinds of losses are adopted to guide the training of PSMM-Net. One kind of loss corresponds to the losses of the three SD-Nets, *i.e.* RGB, Depth and IR modalities data, denoted as $\mathcal{L}^{color}$, $\mathcal{L}^{depth}$ and $\mathcal{L}^{ir}$, respectively. Another kind of loss bases on the summed features from all SD-Nets and the shared branch, which guides the entire network training, denoted as $\mathcal{L}^{whole}$, which . The overall loss $\mathcal{L}$ of PSMM-Net is denoted as follow:
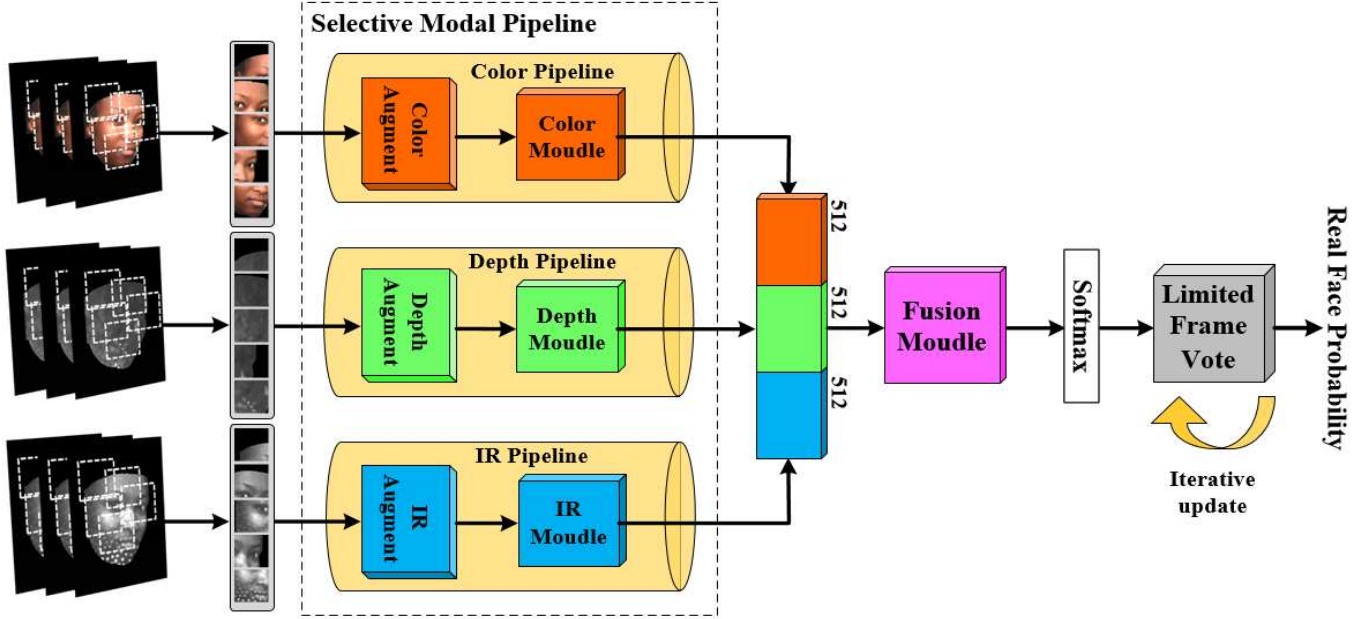
Figure 3. The overall model architecture of PipeNet with Selective Modal Pipeline (SMP) module and Limited Frame Vote (LFV) module. Face patches of three modalities data are fed into corresponding RGB, Depth and IR pipelines in SMP module. The outputs of SMP are concatenated and sent into fusion module for further feature abstraction. LFV is applied to calculate the final output.

$$\mathcal{L} = \mathcal{L}^{whole} + \mathcal{L}^{color} + \mathcal{L}^{depth} + \mathcal{L}^{ir} \qquad (1)$$

4. Methodology

We focus on fusion network for CASIA-SURF CeFA dataset, since fusion network achieved better stability and robustness when facing various kinds of presentation attacks with previous CASIA-SURF dataset.

4.1. The overall Model Architecture

In this work, we propose a novel Pipeline-based CNN fusion architecture for multi-modal face anti-spoofing, named PipeNet. It is based on modified SENet154, with a Selective Modal Pipeline (SMP) module for multiple modalities of image input and a Limited Frame Vote (LFV) module for sequence input of video frames.

Figure 3 shows the overall network architecture. The inputs are facial videos in 3 modalities, which are RGB, depth and IR. For each modal, we take one frame as input and randomly crop it into patches, then send them to the corresponding modal pipeline. In each modal pipeline, data augmentation and CNN feature extraction are performed. The outputs of three pipelines are concatenated into one and sent to the fusion module for further feature abstraction. After linear connection, we obtain the predictions for each frame, and send all of them to the Limited Frame Vote module for iterative calculation. The output is the real face prediction probability of the original facial video.

4.2. Closing the Gap in Cross-ethnicity Learning

As shown in Table 1, in the dataset CASIA SURF CeFA Protocal-4 (including 4@1,4@2,4@3), the faces are from different ethnicities between train and test set, which is called Cross-ethnicity learning. Figure 2 shows the cropped faces of different ethnicities(African, East Asian and Central Asian) in 3 modalities(RGB, depth and IR) in the case of real face. In intuitive observation, the biggest gap between ethnicities is skin tone in RGB modal. There is a potential risk of increasing test loss and ACER error rate, because test data is inconsistent with training data. To eliminate or reduce the gap in the input feature distributions between train and test set, we convert the RGB images of both set into other representations such as HSV, YCbCr and Grayscale, and pick the representation with best performance. The objective is to reduce the influence of skin tone information.

4.3. Selective Modal Pipeline for Different Modalities Data

In essence, face anti-spoofing attack detection is an image classification task. It mainly contains two phases: CNN feature extraction with the first half CNN layers and feature map (binary) classification with other CNN layers.

Because the face image in different modalities con-

Modalities	Crop face	Crop patches	Color trans
RGB	√	√	RGB2Gray
Depth	√	√	×
IR	√	√	×

Table 2. Selected augmentation methods for 3 modalities.

tains specific feature distribution and attributes, we treat them differently with selective pipeline in feature extraction phase. The *pipeline* concept means it is an end-to-end data flow and it can be created and configured in a very flexible way; *selective* means the structure of each pipeline is adapted to each modal. Pipelines for different modalities are independent which could be unified or specific.

We pick up basic block candidates from ResNet[10], ResNeXt[33], XceptionNet[6], SENet[11] *et al.* and we align the input and output dimensions of these blocks to link them as a pipeline and deploy these pipelines to feature extraction module with input data from different modalities. We search for the best combination of three pipelines with experiments. Then, we concatenate the output from SMP as the input of the fusion module. The selective pipeline structure greatly improves the accuracy and efficiency in finding the most suitable structure.

4.4. Limited Frame Vote for Video Classification

Since the input is in video-clip format, it is a sequence of continuous frames with different lengths. Accordingly, the proposed method adds a Limited Frame Vote (LFV) module to obtain the final statistical prediction of a video-clip input.

Several algorithm can provide a measurement of the central tendency of a probability distribution or the random variable characterized by that distribution. The performance mainly depends on the data distribution. Mathematical expectation and geometric median are the most common methods for samples statistics.

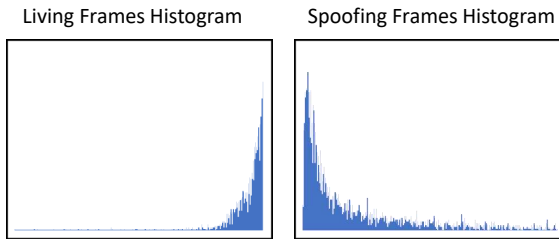Living Frames Histogram   Spoofing Frames Histogram

Figure 4. Frame Probability Distribution.

To avoid outliers, voting should be applied by the partial probabilities of frames. We create the following algorithm based on PauTa Criterion algorithm. Algorithm pseudocode is as Algorithm 1:

In this algorithm, P_i is input, λ and τ are hard value from

Algorithm 1 Limited Frame Vote algorithm

Input: P: list of frame prediction probabilities;
       N: frames length;
       λ: σ offset factor;
       τ: σ offset threshold;

Output: E

1: initial N=length of whole frames in original clip
2: **repeat**
3:    compute expectation $E = \frac{\sum\limits_{i=1}^{N} P_i}{N}$;
4:    compute std $\sigma = \sqrt{\frac{\sum\limits_{i=1}^{N} (P_i - E)^2}{N}}$;
5:    filter P with below equation and create new set to assign to P:

$$P_i = \begin{cases} P_i & P_i \in (E - \lambda * \sigma, E + \lambda * \sigma) \\ 0 & else \end{cases}$$

6:    compute number of Non-zero elements in P and assign to N;
7: **until** $\sigma \leq \tau$

---

experience.

4.5. Some Other Tricks in Training

There are also some training tricks doing benefit to our final accuracy. 4.5.1 is borrowed from winners [28, 40] in Face Anti-spoofing Attack Detection Challenge@CVPR2019 [1]; 4.5.2 and 4.5.3 are breakthrough of previous classic works.

4.5.1 Random Input Crop and Modal Dropout

During training, we apply patch-level input into Face Anti-spoofing area and prove that patch-level inputs can better extract spoof-specific discriminative information over the whole face area. And random dropout of one selective modal is adopted since it can avoid over-fitting and better make use of the characteristics in all three modalities. In the inference phase, we use fixed patches and full modalities to facilitate reproducing result.

4.5.2 Dynamic Start Point for Cosine Decay Restarts

Cyclical cosine annealing learning rate [20] is able to avoid falling into local minimum extreme point during gradient descent process. We improve it by using dynamic learning rate (LR) starting point for each cycle, instead of the fixed value in original version.

Leader	Team	Affiliation	APCER	BPCER	ACER	Rank
Zitong Yu[39, 32]	BOBO	Oulu	1.05±0.62	1.00±0.66	1.02±0.59	1
Zhihua Huang	Super	USTC	0.62±0.43	2.75±1.50	1.68±0.54	2
Qing Yang	Hulking(our team)	Intel	3.25±1.98	1.16±1.12	2.21±1.26	3

Table 3. Top-3 reproduced test set results by organizer in the final stage of Chalearn Multi-modal Cross-ethnicity Face anti-spoofing Recognition Challenge@CVPR2020. The mean and standard deviation values are calculated among results of *Protocol 4*, 4@1, 4@2 and 4@3.

4.5.3 Train/Validation Loss Balance Strategy

Snapshot ensemble [12] inspired us on how to search for best snapshot point of the weight checkpoints during training process. We make a checkpoint-auto-save system whose input are ACER and train/validation loss. It automatically saves checkpoints in multiple positions in each cycle for potential global best.

5. Competition Details

5.1. Dataset and Protocol

We use CASIA-SURF CeFA dataset, which is introduced in Section 3. In order to promote the competition and increase the level of challenge, the most challenging protocol *Protocol 4* designed via combining the condition of both Protocol 1 and 2, was adopted as the evaluation criterion of the competition. As shown in Table 1, it contains 3 data subsets: training, validation and testing sets containing 200, 100, and 200 subjects, respectively. In addition, it has 3 sub-protocols (*i.e.*, 4@1, 4@2 and 4@3) and in which one ethnicity is used for training and validation, and the remaining two ethnicities are used for testing. Moreover, the factor of PAIs are also considered in this protocol by setting different attack types during the training and testing phases.

5.2. Evaluation Metrics

From the evaluation of prediction results, we can get True Positive(TP), False Positive(FP), True Negative(TN), False Negative(FN). Based on these four values, we can calculate the following evaluation metrics. Donate Attack Presentation Classification Error Rate as:

$$APCER = FP/(TN + FP) \qquad (2)$$

Donate Bona Fide Presentation Classification Error Rate:

$$BPCER = FN/(FN + TP) \qquad (3)$$

Average Classification Error Rate is donated as:

$$ACER = (APCER + BPCER)/2 \qquad (4)$$

Eventually, the competition uses the ACER to determine the final ranking.

5.3. Training Configuration

We divided available data into train/validation set with 15:1 ratio with more detailed description in Section 6.1. The optimization method is SGDR and learning rate is set to decay from 0.1 to 0.001. We train the model with 5 cycles, each of which contains 100 epochs.

5.4. Competition Results

We won the 3rd place in the final ranking based on competition organizer's reproduced results. The final result of team scores and ranking is shown as Table 3. In the Chalearn Multi-modal Cross-ethnicity Face Anti-spoofing Recognition Challenge@CVPR2020 [17], our final submission gets the score of 1.97 ± 1.10 ACER on the test set, while reproduced result is 2.21 ± 1.26 ACER, very close to our submission.

Protocal	APCER	BPCER	ACER
4@1	1.44	0.00	0.72
4@2	4.11	1.50	2.80
4@3	2.55	2.24	2.40

Table 4. Best Results of Three Models Corresponding to Protocol 4@1, 4@2 and 4@3.

Table 4 shows the best scores of our submission for 3 trained models corresponding to 4@1,4@2 and 4@3 protocols, which are 0.7222, 2.8055 and 2.4027, respectively.

PipeNet results in Table 3 (Hulking) and Table 4 are generated with same configurations.

6. Experiments and Results Analysis

6.1. Effect of Different Train/Validation Segmentation Ratio

How to segment the training and validation dataset also affects the final result. Several ratios have been tried, including 15:1,12:1,10:1 and 3:1. We ultimately choose 15:1 which generates best performance among them. The test set is much larger than training and validation set and includes many non-existing presentation attacks types in training set. It is observed that a larger training set may still improve the model to learn more information about the unseen samples in test set.

6.2. Effect of Selective Modal Pipeline

As shown in Table 5, we have 5 proposals for pipeline portfolio, including 4 unified designs and 1 selective design. *Unified* means pipelines for RGB, depth and IR are exactly the same while *selected* means different. We examine their ACER on dev set and test set simultaneously. In particular, we adjust the augmentation parameters and CNN blocks combination for each of the three pipelines. CNN blocks consist of SE-ResNetBottleneck and 3 types of SE-ResNeXtBottleneck. The result shows that ACER is reduced after treating each model independently and customizing each modality with different pipeline.

For this section, we repeat each quantitative experiment three times for each of the three dataset protocols. The experiments include training and inference phases. We use the average value as the final result to ensure stability and consistency of model performance.

Pf.	Pipelines			ACER
	RGB	Depth	IR	
1	SRB	SRB	SRB	5.76±1.72
2	SRXB22	SRXB22	SRXB22	4.17±2.09
3	SRXB24	SRXB24	SRXB24	5.31±0.62
4	SRXB34	SRXB34	SRXB34	10.34±9.28
5*	SRXB22	SRB	SRXB22	1.97±1.10

Table 5. Results of Different Pipelines Fusion Portfolios. Portfolio 1-4 contain the same pipeline for RGB, depth and IR modalities, while Portfolio 5 customizes each modality with different pipeline. *SRB* is short for SE-ResNetBottleneck; *SRXB34* is short for SE-ResNeXtBottleneck, layer1 and layer2 repeat 3 times and 4 times (result of portfolio 5 is final submission score, better than the reproduced one in Table 3).

6.3. Does Frame Time Sequence Order Matter?

As shown in Table 6, before the decision of adopting LFV module, we perform experiments to measure the effect of sequence order on video classification results. We take part of continuous frames from sample video as our input, which is corresponding to micro face motions or static face frames, from both living and spoofing video samples, respectively. We then modified PipeNet data augmentation module to process continuous frames. We rearrange the order of frame sequence and test it with three different orders - original order, reverse order and random order. N/A are the cases where dataset does not contain corresponding examples.

The probabilities of predictions among different order options are very close. Two possible reasons are: a) the order of frame sequence is not a key contribution factor in fusion network compared to RGB single-modal network; b) Even after basic cleaning and facial data alignment, the

Protocol	Order	Probabilities		
		Static	Smile	Blink
Living	Original	0.9999	0.9997	0.9999
	Reverse	0.9999	0.9998	0.9999
	Disorder	0.9999	0.9999	0.9999
Print Photo	Original	5.22e-6	N/A	N/A
	Reverse	3.67e-6	N/A	N/A
	Disorder	3.00e-6	N/A	N/A
Video Replay	Original	6.74e-5	N/A	0.0005
	Reverse	5.06e-5	N/A	0.0002
	Disorder	5.64e-5	N/A	0.0002
3D Print Mask	Original	0.0968	N/A	N/A
	Reverse	0.1058	N/A	N/A
	Disorder	0.0935	N/A	N/A

Table 6. Experiments on Effect of Input Frames' Order. N/A are the cases where dataset does not contain corresponding examples.

dataset frames are still not consistent and have various artifacts such as random transportation and partially cropped from background disturbs the context information.

As a result, we choose Limited Frames Vote strategy to obtain the probability of real face for each video sample.

7. Conclusion

In this paper, we propose a flexible and practical multi-stream network architecture to build a robust face anti-spoofing system. The model, named PipeNet, is a pipeline-based design with Selective Modal Pipeline module and Limited Frame Vote module. The quantitative experiments show that the customized pipeline for each modality in PipeNet can make better use of different modalities data. In addition, LFV module provides stable and accurate prediction with continuous frames input. We apply the proposed PipeNet to Chalearn Multi-modal Cross-ethnicity Face Anti-spoofing Recognition Challenge@CVPR2020 [17] and win the third place. Our final submission achieves a score of 2.21 ± 1.26 ACER on the test set. Our future research will focus on enhancing the self-adjustment capability of modal pipelines.

References

[1] *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019.

[2] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 319–328. IEEE, 2017.

[3] Serge Bobbia, Yannick Benezeth, and Julien Dubois. Remote photoplethysmography based on implicit living skin

tissue segmentation. In *2016 23rd ICPR*, pages 361–365. IEEE, 2016.

[4] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, page 1–1, 2016.

[5] I. Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. *BIOSIG*, pages 1–7, 01 2012.

[6] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[7] Tiago de Freitas Pereira, Andre Anjos, Jose Mario De Martino, and Sebastien Marcel. Can face anti-spoofing countermeasures work in a real world scenario? *2013 International Conference on Biometrics (ICB)*, Jun 2013.

[8] Litong Feng, Lai-Man Po, Yuming Li, Xuyuan Xu, Fang Yuan, Terence Chun-Ho Cheung, and Kwok-Wai Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *Journal of Visual Communication and Image Representation*, 38:451–460, Jul 2016.

[9] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. Rank pooling for action recognition. *TPAMI*, 39(4):773–787, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[12] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free, 2017.

[13] Wonjun Kim, Sungjoo Suh, and Jae Joon Han. Face liveness detection from a single image via diffusion speed model. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 24, 04 2015.

[14] Jukka Komulainen, Abdenour Hadid, and Matti Pietikainen. Face spoofing detection from single images using texture and local shape analysis. *Biometrics, IET*, 1:3–10, 03 2012.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[16] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. *2016 6th IPTA*, Dec 2016.

[17] Ajian Liu, Xuan Li, Jun Wan, Sergio Escalera, Escalante Hugo Jair, Meysam Madadi, Zhuoyuan Wu, Xiaogang Yu, Zichang Tan, Qi Yuan, Ruikun Yang, Benjia Zhou, Guodong Guo, and Stan Z Li. Cross-ethnicity face anti-spoofing recognition challenge: A review. *arXiv*, 2020.

[18] Ajian Liu, Zichang Tan, Xuan Li, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z. Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing, 2020.

[19] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *European Conference on Computer Vision*, pages 85–100. Springer, 2016.

[20] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2016.

[21] Jukka Maatta, Abdenour Hadid, and Matti Pietikainen. Face spoofing detection from single images using micro-texture analysis. *2011 International Joint Conference on Biometrics (IJCB)*, Oct 2011.

[22] Ewa Magdalena Nowara, Ashutosh Sabharwal, and Ashok Veeraraghavan. Ppgsecure: Biometric presentation attack detection using photopletysmograms. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 56–62. IEEE, 2017.

[23] Timo Ojala, Matti Pietikäinen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24:971–987, 08 2002.

[24] Gang Pan, Lin Sun, Z. Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. pages 1–8, 01 2007.

[25] Aleksandr Parkin and Oleg Grinchuk. Recognizing multimodal face spoofing with face recognition networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[26] Keyurkumar Patel, Hu Han, and Anil Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security*, 11, 06 2016.

[27] Keyurkumar Patel, Hu Han, and Anil K. Jain. Cross-database face antispoofing with robust feature representation. *Lecture Notes in Computer Science*, page 611–619, 2016.

[28] Tao Shen and Yuyu Huang. Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. 08 2019.

[29] Lin Sun, Gang Pan, Zhaohui Wu, and Shihong Lao. Blinking-based live face detection using conditional random fields. In Seong-Whan Lee and Stan Z. Li, editors, *Advances in Biometrics*, pages 252–260, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[30] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *European Conference on Computer Vision*, pages 504–517. Springer, 2010.

[31] Jue Wang, Anoop Cherian, and Fatih Porikli. Ordered pooling of optical flow sequences for action recognition. In *WACV*, pages 168–176. IEEE, 2017.

[32] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *CVPR*, 2020.

[33] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[34] Zhenqi Xu, Shan Li, and Weihong Deng. Learning temporal features using lstm-cnn architecture for face anti-spoofing. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 141–145, 2015.

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

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

[37] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Li. Face liveness detection with component dependent descriptor. pages 1–6, 06 2013.

[38] Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifeng Li, and Wenqiang Liu. Face anti-spoofing: Model matters, so does data. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3502–3511, 2019.

[39] 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*, 2020.

[40] Peng Zhang, Fuhao Zou, Zhiwen Wu, Nengli Dai, Skarpness Mark, Michael Fu, Juan Zhao, and Kai Li. Feathernets: Convolutional neural networks as light as feather for face anti-spoofing, 2019.

[41] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guogong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z. Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, page 1–1, 2020.