Developing a Standard for Antispoofing Using Multiple Frames Approach: A Baseline

Hanry Ham[1], Ari Saptawijaya[1], and Aniati Murni Arymurthy[1]

Faculty of Computer Science, Universitas Indonesia
hanry.ham@ui.ac.id, {saptawijaya, aniati}@cs.ui.ac.id

Abstract. Face recognition systems are becoming more susceptible to spoofing attacks, where an attacker employs a counterfeit face to mimic an authorized person. Multiframe-based techniques have demonstrated promising outcomes in improving supplementary tasks, such as creating more robust depth maps compared to single-frame approaches. Nevertheless, the main focus in the state-of-the-art (SOTA) remains on single-frame usage. At present, there are no a standardized baseline methods available, even for basic binary classification, particularly targeting Domain Generalization, specifically the OCIM protocol. Numerous previous studies have utilized their unique evaluation methods, which often lack clarity and detailed information on frame sampling in multi-frame formats, making reproducibility challenging. This work aims to tackle these issues by setting a standard for face antispoofing (FAS) techniques using multiple frames. The proposed method will create and evaluate the generalization of antispoofing system based on multiple frames. The research objective include: (1) Develop a foundational baseline for utilizing multiple frames in FAS; (2) Enhance the generalization capability of the FAS model by leveraging multiple frames. The anticipated results of this study encompass a robust multi-frame-based face antispoofing method that can enhance generalization in intertest evaluations, along with an understanding of the constraints and possible advancements of the proposed framework.

Keywords: Multiple frames FAS · Domain Generalization Multiple Frames FAS · Baseline Multiple Frames FAS

1 Introduction

Face antispoofing (FAS) stands as a crucial element in Biometric Verification technology, working in tandem with Face Recognition (FR). The purpose of face antispoofing is to thwart attempts at spoofing faces using physical attacks. The commercial industry has embraced biometric verification under the banner of Know Your Customer (KYC), which incorporates both Face Recognition and face antispoofing. As identity attacks in FR evolve and become more sophisticated, the threat of fake identity attacks designed to deceive the system increases. This could potentially result in fraudulent transactions across various industries, leading to substantial financial losses.

Face antispoofing, also referred to as face liveness in some studies, is essentially a risk to an individual's identity. This identity could be manipulated or falsified by printing someone's face using various types of printers and paper. Attacks could be carried out not only through printed mediums, but also using screens. The resolution of these screens can range from commonly available ones to advanced OLED screens. Besides these printed and screen-based attacks, the advanced one could also be made in the form of realistic face masks. The primary goal of face antispoofing System (FAS) is to counteract these types of attacks.

A key research area in FAS that many state-of-the-art models are attempting to address is the model's generalization. Prior studies indicate that the face antispoofing issue extends beyond a simple binary classification, presenting itself as a fine-grained problem [12, 17, 23]. The term 'Fine-Grained classification' is used to describe the minor variations in training data that is not exist in test data, such as spoof type attacks, different types of medium spoof, and environmental factors, which could potentially result in erroneous inference outcomes. Consequently, Domain Generalization (DG) has emerged as a prominent area of research in FAS. DG is an approach where the training domain does not have access to the test domain, with the goal of assessing the model's generalization robustness.

The progression of FAS has been characterized by the onset of binary classification [22], the implementation of depth networks with extra supervision [1], the combination of depth and rPPG signals [10], the breakdown of spoof traces into spoof noise and genuine faces [8], and the employment of generative AI [26] to facilitate the network's learning from synthetic datasets without labels, in an unsupervised manner. In addition, the leading methods exhibited the application of metric learning [7, 9, 14] which resulted in a significant performance jump compared to the other method. Recently, the FAS issue has also seen the application of LLM techniques, following the launch of FLIP [13], a framework that employs CLIP [11] as the underlying pretrained model and is presently at the forefront of evaluation metrics using the DG method.

The methods mentioned above all utilize the same input approach, which is a single frame, as is common in general image classification as well as in Face Anti Spoofing. The objective of this research is to investigate the application of multiple or temporal frames as input. The exploration of multiple frames is based solely on the fundamental concept that FAS is a fine-grained classification, and the use of multiple frames could enable the network to learn more and extract finer details compared to the single frame approach.

The concept of employing temporal frames is not a novel proposition in this study. Wang et al. [18] highlight that the depth map produced by the auxiliary network using temporal frames yields a superior depth map relative to the one generated using a single frame. This leads to the assurance of a robust generalization. Furthermore, the initial temporal approach in FAS is discovered in Xu et al. [21], where the research employed the pre-established CNN architecture stack with an LSTM unit for every frame. Subsequently, a year later, Gan et al. [5] delved into the first application of 3D CNN with AlexNet to fully harness the

temporal frames. These methods have only been evaluated on casia and casia-replay attack subsequently, with the evaluation being conducted to specifically assess the intra-test. The recent approach known as Geometry-Aware Interaction Network (GAIN) [3] has pioneered the integration of spatial and temporal elements, employing dense face landmarks to extract additional features that differentiate authentic and spoof facial movements.

Since then, the exploration of multiple frames in face antispoofing does not evolve that fast compared to the number of methods published under single frame approaches. The current SOTA in utilizing the multiple frames are exploring the exploitation of more robust proposed auxiliary network [3, 18]. As this research would like to highlight the Domain Generalization capability in FAS.

To the author's knowledge, no current methods have explored domain characteristics using multiple frames up to now. There are no baseline comparisons with consistent settings in a multi-frame format, and previous studies lack details on sampling frames for training or on evaluation processes. Each proposed method employs its own evaluation criteria, making it challenging to assess the performance improvements of these methods.

A successful method in the single-frame domain is the metric learning approach, which leverages the intrinsic characteristics of each training dataset's domain to enhance the differentiation between live and spoof labels in their embeddings. Currently, no studies have further explored these intrinsic factors using multiple frames. Implementing such a framework would not only extract more features but also enhance the intrinsic features of each domain dataset, aiming for a more generalizable FAS model.

2 Previous Works

As some works explained that FAS is a fine grained classification [18], thus one of the work to be able to extract a robust features is through the temporal frames. There are some advantages of using multiple frames described by Wang et al. [18] work where their temporal works aim at enhancing the depth map generated. It is proved that in their work the Depth map generated using temporal frames are more robust compared to the single frame.

The dual approach to multiple frame analysis can be categorized into two types: one employing Conv2D and the other using Conv3D. The first implementation of Conv2D combined with LSTM was introduced by Xu et al. [21], which was solely assessed using the intertest CASIA dataset. The concept of incorporating multiple frames for input was initially proposed in Gan et al. [5], employing a 3D CNN, with evaluations conducted on the Replay-Attack and CASIA datasets through intra-test. Both studies briefly explored the impact of varying the number of frames used during the training phase.

A significant challenge in developing inputs from multiple frames is the presence of inconsistent predictions within each frame of a video clip, as discussed in [20]. It is assumed that a video clip, whether of a live or spoof subject, is

assigned a consistent label that applies to all its frames. However, predictions tend to vary across different frames.

To address the **temporal inconsistency**, two new loss functions, named Temporal Consistency and Class Consistency Loss, were introduced by Xu et al. [20]. Temporal Consistency Self-Supervision ($\mathcal{L}_t$) aims to maintain temporal consistency across several frames. Class Consistency Loss ($\mathcal{L}_e$) ensures that embeddings from the same class but different videos are similar. The classification head uses the last frame as a reference, the temporal loss considers the last three frames, and the class consistency loss is applied starting with the first frame and continues up to the third-to-last frame. This loss functions are described in equation 1.

$$\mathcal{L}_t = \frac{1}{m}\sum_{i=0}^{m} \underset{i,j \in v}{max} \parallel x_i - x_j \parallel_2^2; \qquad \mathcal{L}_e = \frac{1}{m}\sum_{i=0}^{m} max\ y_{ij} \parallel x_i - x_j \parallel_2^2 \tag{1}$$

where m is the batch size, x_i and x_j represent temporal frames from the same video represents with v. Here, y_{ij} is 1 if x_i and x_j belong to the same class in the batch; otherwise, y_{ij} is 0.

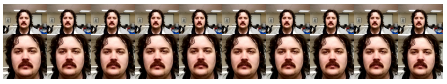
Metric Learning. One of the concept that utilizing all the intrinsic features from datasets such as separating the dataset source or spoof type attacks proved that able to improve the generalization of the model especially in FAS.

Asymmetric Triplet Mining Loss. The loss function proposed by Jia et al. [7] in SSDG was designed to disperse spoofing attacks in the feature space while keeping real ones close together. SSDG noted that spoofing attacks have a larger variety of ways to be captured, so it is beneficial to have a dispersed feature space. In contrast, the discrepancies between real ones are much smaller, so they should be kept close together.

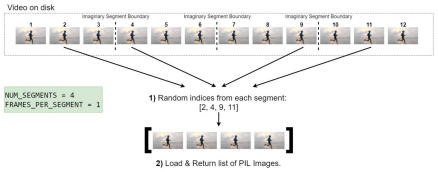
Domain Invariant Concentration Loss. introduced by Liao et al. [9] loss function unifies real faces from all domains into a single group and attempts to learn their features embeddings that are invariant to this group. The real images are forced to fit in the center of the embedding space, while the spoof images are treated the same, despite each domain having its own characteristics of spoof type attacks. This loss classifies the embeddings based on each corresponding spoof type attack and is invariant to the source of the domains. It is also worth noting that the author noticed that the combination of a transformer model so called DiVT and this loss is already powerful in feature learning of the whole face, compared to previous works that used complex adversarial training mechanisms.

Separability and Alignment Loss. The initial concept implementing this technique proposed by Sun et al. [14] in facial anti-spoofing, also known as SA-FAS. Prior methods centered solely on separating embedding features. However, separability by itself is not enough to enhance domain generalization. The isolated feature clusters can reside anywhere within the feature space, causing the domain-specific optimal hyperplane to remain inconsistent. Consequently, the global classifier may still erroneously integrate spurious correlations.

3 Methodology



(a) Preparing datasets visualization



(b) Video Frame Sampling Method used in Temporal Segment Networks for Action Recognition [16]

Fig. 1: The preprocessing and sampling methods in proposed baseline for Multiple Frames based FAS

3.1 Preparing Datasets

Four datasets are used: oulu-npu [2], casia-mfsd [25], idiap replay-attack [4], and msu-mfsd [19]. All datasets are in video format. Utilizing the single frame method, which involved cropping and aligning the face, this project also involved cropping each frame using the MTCNN Face Detector [24] to a dimension of 256x256. This process is described in Figure 1a. The first row represents the extracted video into frames. The second row represents the input frames that used in the experiment after preprocessed by face detector.

3.2 Handling Multiple Frames

The literature review does not provide specific details on the extraction of multiple frames [3, 17, 20, 21]. Consequently, the authors have referenced techniques from action recognition, which similarly utilizes multiple frames. This study adheres to the established methods for frame extraction as described in Wang et al. [16]. The dataloader's approach to managing multiple frames is depicted in Figure 1b.

To fully understand the video frame sampling technique thoroughly, two key variables are introduced as follows:

1. *NUM_SEGMENTS*
   NUM_SEGMENTS refers to a variable that partitions the total number of frames in the video into equal segments.
2. *FRAMES_PER_SEGMENT*
   FRAMES_PER_SEGMENT denotes a variable that samples a specific number of frames within each segment. The selection of frames can either be random or can target the central frame of each segment.

For example, given NUMBER_SEGMENTS = 4 and FRAMES_PER_SEGMENT = 1, the video shown in Figure 1b contains a total of 12 frames. Consequently, 4 segments will be created, each containing 3 frames. Within each segment, only 1 frame will be sampled as specified by FRAMES_PER_SEGMENT. Therefore, a total of 4 frames will be forwarded to the dataloader.

3.3 Pretrained Model

All the feature extractor or image encoders used in this experiments are using ResNet 18 both in using Conv2D and Conv3D. As the convolution is differ then the pretrained model will be different as well. All the Conv2D using pretrained from ImageNet (1000 classes), while the Conv3D using pretrained [6] from combination of Kinetics 700 (700 classes) and Moments in Time (339 classes) where in total has 1030 classes.

4 Baseline Multiple Frames FAS

The multi-frame FAS approach lacks baseline data, especially in cross-domain contexts. Therefore, it is crucial to create a robust foundation for baseline comparisons for both CNN2D and CNN3D. The evaluation metrics are determined at the video level. Video level implies that the total number of frames used during training will be matched with the evaluation data. Consequently, all frames within a single video clip are considered as one video level.

4.1 Temporal - ResNet18

In the baseline configuration, LSTM will be employed alongside Conv2D to fully capture the temporal aspects. The image encoding is performed using ResNet 18. This architecture is illustrated in Figure 2a. The input consists of a video, and the dataloader is structured into five dimensions: Batch B, the number of frames per video (N), Image channels (C), Height (H), and Width (W).

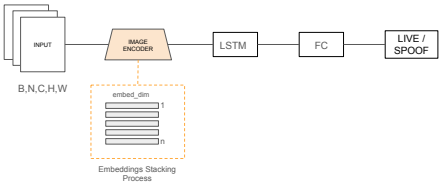
The Conv2D in this context is limited to processing a single frame at a time. As a result, each frame from the batch is individually extracted and subsequently processed to merge the embeddings by stacking them. These stacked embeddings are then fed into an LSTM for further processing. At the output of the LSTM, a fully connected layer is utilized, which makes use of a Cross Entropy loss function.

4.2 Temporal - ResNet3D 18

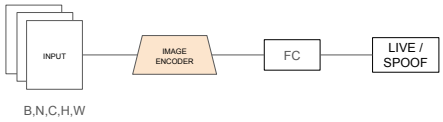
For the baseline Conv3D is also using ResNet3D18 [6]. As Conv3D could handle multiple frames, subsequently the frames directly feed into the fully connected layer with Cross Entropy loss function to predict the live/spoof images. This network is describe in Figure 2b.

4.3 R2D18 FC ENC - LSTM

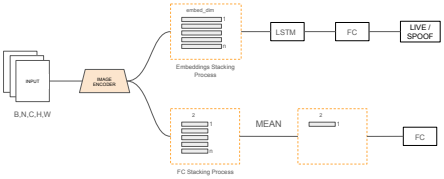
Xu et al. [20] notes that during temporal processing, fluctuations in frame predictions within a video clip can occur. To maintain consistent class predictions across frames, it is necessary to implement temporal class consistency. This is achieved by applying an additional Cross-Entropy loss, which incorporates the average of all FC embeddings. The details of this architecture is described in Figure 2c.



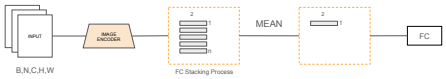
(a) Network Temporal ResNet18



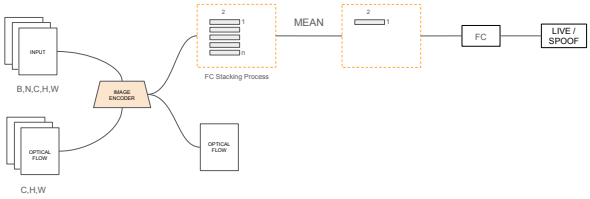
(b) Network Temporal ResNet3D 18



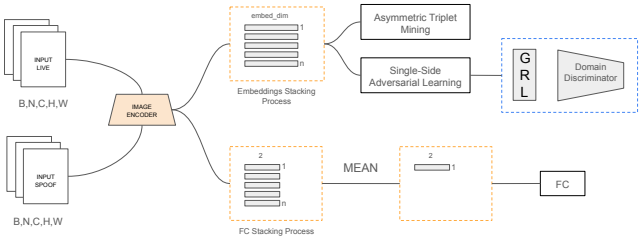
(c) Network Temporal R2D18 with LSTM and Temporal Class-Consistency loss



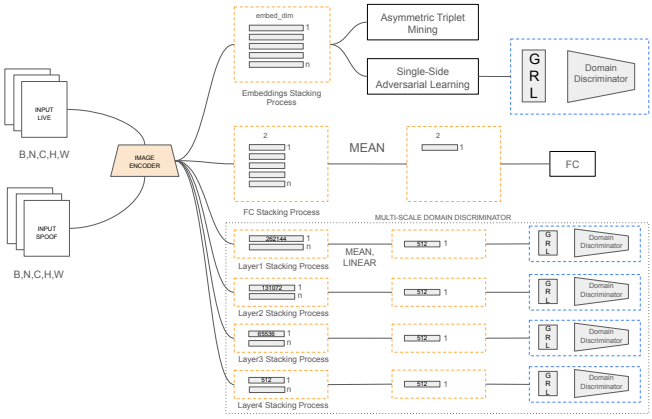
(d) Network Temporal R2D18 with no LSTM and Class-consistency loss



(e) Network Temporal R2D18 with Optical Flow



(f) Network Temporal R2D18 with SSDG loss



(g) Network Temporal R2D18 with Multi-Scale Domain Discriminator

Fig. 2: The proposed baseline methods of the Multiple Frames based FAS

The image encoder generates two outputs: Firstly, stack embeddings of length embed_dim, which are subsequently used for LSTM processing. Secondly, a stack FC tensor of length 2, which is employed in computing the temporal consistency loss by averaging the mentioned tensors in equation 2. All the loss functions employed in this network are detailed in equation 3.

$$\mathcal{L}_{consistency} = -\frac{1}{b}\sum_{i=0}^{n} log(p_{mean_{yi}}) \tag{2}$$

$$\mathcal{L} = \mathcal{L}_{consistency} + \mathcal{L}_{cls} \tag{3}$$

During inference, the embedding stacking process branch is utilized, which passes through the LSTM network. Finally, the FC layer is activated with the softmax function to produce the live/spoof value.

4.4 R2D18 FC ENC - NO -LSTM

This network assesses the influence of temporal effects on the generalization capabilities of face antispoofing methods. It employs a temporal consistency loss function, which incorporates the Cross-Entropy loss as detailed in equation 4. All loss function used in this network are detailed in equation 5. The image encoder used is ResNet 18. The details of this architecture is described in Figure 2d

$$\mathcal{L}_{consistency} = -\frac{1}{b}\sum_{i=0}^{n} log(p_{mean_{yi}}) \tag{4}$$

$$\mathcal{L} = \mathcal{L}_{consistency} \tag{5}$$

4.5 R2D18 - NO - LSTM - Optical Flow

Optical flow can only be accomplished when multiple frames are used as input. Currently, there are no methods that employ optical flow for the DG approach in face antispoofing. Based on previous work suggesting that incorporating additional tasks could enhance model generalization, this study aims to evaluate the impact of this additional task. The optical flow in this context utilizes a pretrained model from RAFT [15].

Figure 2e illustrates the specifics of the proposed network. The image encoder will feature two branches, with the FC stacking process dedicated to the temporal class consistency loss describe in equation 6. The generated optical flow is upscaled to 256 and subsequently trained using MSE loss to learn the constructed optical flow, it describe in equation 7. All training loss is describe in equation 8.

$$\mathcal{L}_{consistency} = -\frac{1}{b}\sum_{i=0}^{n} log(p_{mean_{yi}}) \tag{6}$$

$$\mathcal{L}_{flow} = (gt_{flow} - y_{flow})^2 \tag{7}$$

$$\mathcal{L} = \mathcal{L}_{consistency} + \mathcal{L}_{flow} \tag{8}$$

4.6 R2D18 - SSDG

Leveraging the SSDG framework [7], which employs asymmetric triplet mining to bring live embeddings closer together while differentiating spoof embeddings based on their source domain, alongside single-sided adversarial learning to enhance discrimination of the live domain. Both of these functions were implemented using asymmetric triplet loss and adversarial loss. The GRL is set the same with the default formula, described in Chapter 2. The details of the architecture is described in Figure 2f

To emphasize, the input divides each dataset into separate domains, each encompassing both authentic and counterfeit images. The image encoder employs the same weights for both types of images. It generates two types of outputs: stack embeddings with embed_dim dimensions and a stack FC tensor with 2 dimensions. Additionally, the stack embeddings are processed by averaging and then supervised using two techniques: asymmetric triplet mining and unilateral adversarial learning. Simultaneously, the average of the stack FC embeddings is calculated and monitored using temporal class consistency loss via Cross Entropy loss. All the loss functions used are describe in equation 9.

$$\mathcal{L}_{SSDG} = \mathcal{L}_{Cls} + \lambda_1 \mathcal{L}_{Ada} + \lambda_2 \mathcal{L}_{AsTrip} \tag{9}$$

4.7 R2D18 - DSDG

This network is still based on SSDG, however there are additional loss functions used called multi-scale domain discriminator to allow the early layer embeddings (layer 1 - layer 4) to be able to distinguish the domain. Domain here means the source of dataset, in the intertest evaluation, there are 3 domains. In order to learn these domain CE loss is employeed. The GRL is set the same with the default formula, described in Chapter 2. The details of the architecture is described in Figure 2g. This network is setup due to our hypothesis that by utilizing the multiple frames in earlier layer might be able to boost the generalization of the model. All the loss functions used are describe in equation 10.

$$\mathcal{L}_{SSDG} = \mathcal{L}_{Cls} + \lambda_1 \mathcal{L}_{Ada} + \lambda_2 \mathcal{L}_{AsTrip} + \lambda_3 \mathcal{L}_{Ada_{layer1}} + \lambda_3 \mathcal{L}_{Ada_{layer2}} + \lambda_3 \mathcal{L}_{Ada_{layer3}} + \lambda_3 \mathcal{L}_{Ada_{layer4}} \tag{10}$$

5 Experiments

The training process utilizes the PyTorch library as the primary Deep Learning framework. Training and evaluation are conducted on a three-GPU setup, consisting of a single RTX 4090 and two RTX 3090 GPUs. Certain experiments with higher frame numbers demanded more VRAM for computation. However, some experiments were still manageable on a single GPU during the training phase.

5.1 Experiments Setup

Datasets and Protocols. As for the experiment, the aim for this work is evaluating the Generalization capability through the Domain Generalization approach

where the approach is intra-test protocol. The 'leave-one-domain-out' approach in the intertest serves as a conventional DG assessment in FAS, designed to test the trained model on an unfamiliar domain/dataset. Four datasets are typically employed, including OCIM.

Implementation Details. Two types of tests were conducted in this study: intra-test and inter-test. The study implements seven multi-frame-based methods as detailed in Section 4, along with one spatial method employing ResNet2D 18 for comparative purposes. Various parameters such as num_segments = {2, 5, 10, 20}, frames_per_segment = $\{1, 2\}$, optimizer = $\{AdamW, SGD\}$, learning_rate = $\{1e^{-4}, 1e^{-3}, 1e^{-2}\}$, and the length of embeddings in image encoder = $\{256, 512, 1024\}$ and the last layer in LSTM before FC layer has embeddings variations of $\{64, 128, 512\}$ were adjusted to determine the optimal configurations for these methods. A single-frame level, labeled as "basic" in the experiment name, is a frequently used single-frame level method that employs Resnet18 with CE loss, adhering to the same evaluation protocol as outlined in [7]. This approach was included in the experiments to compare single-frame and multiple-frame methods. All models underwent training for 40 epochs, after which the HTER and AUC metrics were evaluated to identify the best model for each training process.

Evaluation Details. The frames analyzed in this work are selected at the video level. The number of frames utilized in the evaluation matches the quantity used during training, following an identical sampling protocol. When assessing the influence of the number of frames through hyperparameter tuning, the number of videos evaluated remains consistent to ensure fair comparisons, despite variations in the number of frames in the same training protocol.

Results. The findings are presented in two training configurations: Intra-test and Inter-test. The Intra-test results are shown in Table 1. Similarly, the Inter-test results are illustrated in Table 2. Through all the baselines performed in temporal configurations, the R2D18 is having the worst evaluation metrics, these reflected in all inter and intratest protocols. Meanwhile the R3D18 perform almost 2x times better in comparison with R2D18 for the intratest and intertest protocols. However, simply adopting the multi-frame approach does not ensure optimal outcomes. The used of CE loss at the end of LSTM unit shows an inferior performance. One of the possible reason is that, there is a need for a supervised loss before the stacking embedding further process through the LSTM unit. This underscores the findings of Xu et al. [20], which indicated that multiple frames experience prediction inconsistencies between video clip frames. Introducing a Temporal consistency framework using CE loss could potentially boost the model's performance.

Moreover, to integrate this temporal class consistency loss, it is applied by stacking the fully connected embeddings output for each frame, supervised by the CE loss. This approach is designed to mitigate inconsistency across frames within the same video. For intratest, this configuration is under R2D_FC_ENC and shows significant improvement compared to R2D, as well as a slight superiority over R3D. For intertest, it also demonstrates improvement in almost all

protocols, except for OMI to C, which shows a disparity of around 3.033% for HTER. A notable performance is observed in ICM to O, making this configuration the best among all configurations tested in this study with HTER of 16.819%. The use of temporal consistency loss is crucial in the utilization of temporal frames at least with the network with the basic configurations such as R2D.

This also piques our interest: if the temporal consistency loss is so crucial, it might be feasible to eliminate the LSTM branch with the CE. This setup is detailed under R2D18_NO_LSTM in Figure 2d. During the intratest phase, the results reveal a slight performance degradation on the Oulu test, with a disparity of 0.06% in HTER performance. However, in intertest protocols, this configuration outperforms with an overall average HTER of 20.099%, compared to 21.732% for R2D_FC_ENC.

Furthermore, the intratest results demonstrate excellent performance, as most of the proposed methods already achieve an HTER of 0% and an AUC of 100% in the CIM protocol. For the O protocols, the results show an HTER of $<1\%$ and an AUC $> 99\%$, indicating that intratests in multiple frame protocols are no longer challenging. Thus, for subsequent experiments, no intratest protocols will be used for the remaining proposed network configurations.

As all the basic foundational model has done in the earlier experiments above, this work also would like to utilize the auxiliary network with the utilization of optical flow. As in the single frame, it has the disadvantage of utilizing this as optical flow, due to the needs of multiple frames acquired in order to generate the optical flow. Therefore, this work would like create a baseline of using optical flow combined with R2D_NO_LSTM with experiment name so called R2D_NO_LSTM_OPTFLOW described in Figure 2e. Notably, in the OCI to M scenario, this method achieves the second-best performance with an HTER of 7.835%, trailing the best performance by 0.338, which is held by R2D18 - FC ENC - NO LSTM. Similarly, in ICM to O, it also secures the second-best position with an HTER of 17.529%, lagging behind the top performance by 0.709, achieved by R2D18 - FC ENC. However, in the OMI to C and OCM to I cases, this method ranks second to last with HTERs of 33.438% and 29.045%, respectively.

The additional task of optical flow can only be utilized with the multiple frames approach to produce the optical flow map. The only prior work identified using the implementation of optical flow is by Chang et al. [3]. However, their ablation study lacks detailed information on the evaluation performed and the architecture used. The only known details are that they used 3DCNN + optical flow and evaluated solely on OCM to I, achieving 92.28% in AUC. In the current work depicted in Table 2, the best HTER for OCM to I using optical flow is 29.045% and the AUC is 68.408%. There remains a significant performance gap that this work cannot bridge.

In single frame approach, the used of Metric Learning where in the earlier adopter is SSDG [7]. This experiments also would like to emphasize whether this method if its bring to temporal frame will benefit. The SSDG is selected to inte-

Table 1: Intratest Protocols

exp name	O		C		I		M	
	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
BASIC	3.576	99.529	***3.336***	***99.669***	***0.563***	***99.973***	***6.250***	***98.411***
R3D	0.952	99.873	**0**	**100**	**0**	**100**	**0**	**100**
R2D	8.680	97.130	1.659	99.401	9.305	96.863	4.610	97.752
R2D_FC_ENC	**0.85**	**99.93**	**0**	**100**	**0**	**100**	**0**	**100**
R2D_NO_LSTM	0.91	99.92	**0**	**100**	**0**	**100**	**0**	**100**

Table 2: Intertest Protocols

exp name	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O	
	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)	HTER (%)	AUC (%)
BASIC	23.542	83.453	31.603	72.809	23.875	79.794	*22.083*	*84.286*
R3D	14.649	92.950	28.512	70.593	26.481	71.936	27.245	77.176
R2D	29.298	70.217	42.812	56.716	41.342	57.069	33.870	72.055
R2D_FC_ENC	14.304	90.907	31.545	74.196	24.260	74.411	**16.819**	**89.551**
R2D_NO_LSTM	**7.497**	**96.654**	26.262	70.011	25.787	72.030	20.850	86.672
R2D_NO_LSTM_OPTFLOW	7.835	96.659	33.438	61.418	29.045	68.408	17.529	88.489
R2D_SSDG	8.028	96.442	**21.818**	**84.877**	**10.744**	**95.344**	20.965	85.184
R2D_DSDG_ADVALL_EMB	11.904	94.474	32.456	68.788	28.347	74.908	21.042	83.904

grate to our baseline due to its seperation in domain label only using a different dataset, denoted with the experiment name R2D_SSDG. The SSDG approach with a single frame significantly boosts the model's generalization performance across all evaluations. However, directly applying SSDG with multiple frames fails to achieve the same level of significant performance as the single frame usage. It is important to note that the implementation of R2D_SSDG also incorporates temporal class consistency loss, making this method comparable to R2D_FC_ENC and R2D_NO_LSTM. However, Table 2 indicates that OCI to M falls short of R2D_FC_ENC by a margin of 0.783%. Additionally, ICM to I also lags behind both R2D_FC_ENC and R2D_NO_LSTM by 4.145% and 0.115%, respectively. Although the R2D_SSDG has demonstrated the best overall performance to date, the HTER metrics for 2 protocols still surpass 20%.

As we start from the hypothesis about the utilization of multi scale domain discriminator applied to the R2D_SSDG, within the experiments under R2D_DSDG. The findings indicate that utilizing an earlier layer for the domain discriminator, even with multiple frames, hinders the network's learning process. One of the outcomes demonstrates a tight performance between SSDG and DSDG on ICM to O with HTER of 21.042% loose to SSDG with HTER of 20.965. However, in other inter-test evaluations like OCI to M, it results in an HTER of 11.904%, with a gap of 4.407% from the best. Additionally, for OMI to C and OCM to I, the HTERs are 32.456% and 28.347%, respectively, with both evaluations showing a substantial performance gap of over 10% in HTER.

This study indicates that further examination is required to understand how the embeddings of images react to the proposed loss function and the aggregation method. In this work, the mean of all embeddings from frames extracted from the same video is used.

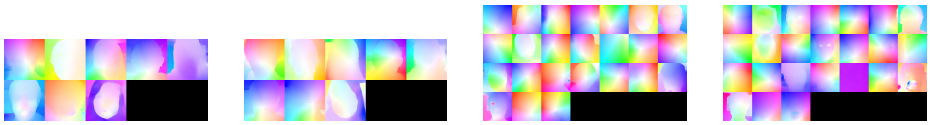
5.2 The impact of total frames used during the training

All experiments were conducted extensively on both intratest and intertest. all experiments were also tested with various values of num_segments and frames_per_segment along with the other hyper-parameters mentioned in implementation details section. The results from both intertest and intratest did not reveal any discernible pattern, suggesting that the number of frames is merely a hyperparameter that may lead to different performance outcomes in various evaluation tests. Through all experiments conducted, do not have a direct correlation with the generalization of the FAS model.

Nevertheless, the quantity of frames contributes to supplementary tasks like producing a reliable depth, as noted in the previous work [18], in contrast to the single frame approach. Typically, an additional task is represented by a new branch in the network, which in this study employs the optical flow task. Prior research has demonstrated that incorporating supplementary tasks can enhance the model's generalization.

5.3 The computed Optical Flow

The quantity of frames utilized to produce the optical flow with the RAFT pretrained model [15] significantly affects the quality of the ground truth. For better view, the generated Optical flow is taken from the smallest number of videos from MSU data shows in Figure 3 display snippets of the generated optical flow from the training data for each dataset. Increasing the number of frames used to generate the optical flow results in a more detailed representation of movement in the optical flow map. Nevertheless, the optical flow results are less promising when compared to SSDG and DSDG. The qualitative distinction between live and spoof images is unclear, indicating that the generated optical flow maps exhibit a blend of live and spoof characteristics within each label.

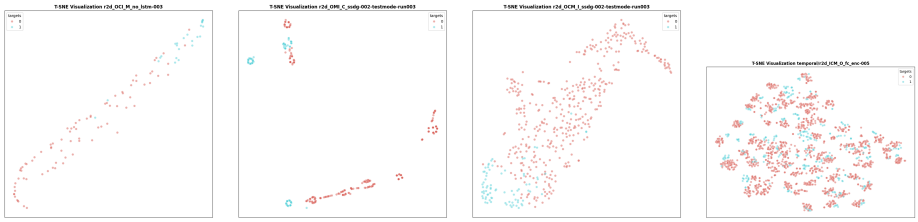


(a) Live MSU - total frames used 10 **(b)** Live MSU - total frames used 20 **(c)** Spoof MSU - total frames used 10 **(d)** Spoof MSU - total frames used 20

Fig. 3: Optical flow generated for Train dataset - MSU

5.4 Visualization of the best Intertest Protocols

This qualitative approach centers on inter-test evaluation, which is the primary focus of this study. The T-SNE method is employed in this research. This technique, commonly used in prior studies, is designed to visualize high-dimensional data. In this study, the T-SNE was performed using the built-in function from the Scikit Learn library.



(a) best OCI to M : R2D_NO_LSTM **(b)** best OMI to C : R2D_SSDG **(c)** Best OCM to I : R2D_SSDG **(d)** Best ICM to O : R2D_FC_ENC

Fig. 4: TSNE of the best HTER evaluation metric in overall intertest protocols in correlation with the best results in Table 2

Each point in the embedding space represents a video, with the extracted frames used during evaluation being consistent with the training configuration. All extracted frames were processed using the mean operator. The mean embeddings from all extracted frames within a video are used as input to the TSNE. The layer chosen to output these embeddings is located one layer before the final fully-connected layer. All visualizations below use the optimal training configurations from each model to show how the trained model extracted features from the test dataset. The visualization includes two labels: 0 for spoof and 1 for real instance.

6 Conclusion

In summary, this work seeks to explore the advantages of employing multiple frames over a single frame and to enhance the generalization of FAS models using a multiple frame approach. These investigations involved establishing a baseline due to the absence of existing baseline in the literature, as well as examining the application of Temporal class consistency, Metric Learning with SSDG loss, and SSDG with a multi-scale domain discriminator.

This study is significant due to its potential to enhance the DG approach employing a multiple frame strategy, focusing on the effects of metric learning and supplementary tasks to boost model generalization. To accomplish this, this work adopts a metric learning method that has not been thoroughly investigated in the context of multiple frames, specifically in terms of how to combine all the embedding from the relevant frames within a video clip.

The anticipated results encompass the methodology for effectively using multiple frames, which will enhance our understanding of the influence of temporal frames on improving the generalization of the FAS model. In summary, this proposed research has the potential to significantly advance the understanding of face antispoofing, and it is convinced that it is crucial to undertake this study to tackle the urgent challenges in Domain Generalization for face antispoofing.

Bibliography

[1] Atoum, Y., Liu, Y., Jourabloo, A., Liu, X.: Face anti-spoofing using patch and depth-based cnns (2017) 2

[2] Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: Oulu-npu: A mobile face presentation attack database with real-world variations. pp. 612–618, IEEE (5 2017), ISBN 978-1-5090-4023-0, https://doi.org/10.1109/FG.2017.77, URL http://ieeexplore.ieee.org/document/7961798/ 5

[3] Chang, C.J., Lee, Y.C., Yao, S.H., Chen, M.H., Wang, C.Y., Lai, S.H., Pei-Chun Chen, T.: A Closer Look at Geometric Temporal Dynamics for Face Anti-Spoofing. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 2023-June, pp. 1081–1091 (2023), ISBN 9798350302493, ISSN 21607516, https://doi.org/10.1109/CVPRW59228.2023.00115 3, 5, 11

[4] Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing (2012), URL http://www.idiap.ch/dataset/replayattack 5

[5] Gan, J., Li, S., Zhai, Y., Liu, C.: 3D Convolutional Neural Network Based on Face Anti-spoofing. Proceedings - 2017 2nd International Conference on Multimedia and Image Processing, ICMIP 2017 **2017-Janua**(1), 1–5 (2017), https://doi.org/10.1109/ICMIP.2017.9 2, 3

[6] Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6546–6555 (2018) 6

[7] Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing (4 2020), URL http://arxiv.org/abs/2004.14043 2, 4, 9, 10, 11

[8] Jourabloo, A., Liu, Y., Liu, X.: Face de-spoofing: Anti-spoofing via noise modeling. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11217 LNCS**, 297–315 (2018), ISSN 16113349, https://doi.org/10.1007/978-3-030-01261-8_18 2

[9] Liao, C.H., Chen, W.C., Liu, H.T., Yeh, Y.R., Hu, M.C., Chen, C.S.: Domain invariant vision transformer learning for face anti-spoofing (2023) 2, 4

[10] Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision (3 2018), URL http://arxiv.org/abs/1803.11097 2

[11] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2 2021), URL http://arxiv.org/abs/2103.00020 2

[12] Shao, R., Lan, X., Yuen, P.C.: Regularized fine-grained meta face anti-spoofing. AAAI 2020 - 34th AAAI Conference on Artificial Intelligence pp. 11974–11981 (2020), ISSN 2159-5399, https://doi.org/10.1609/aaai.v34i07.6873 2

[13] Srivatsan, K., Naseer, M., Nandakumar, K.: Flip: Cross-domain face anti-spoofing with language guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 19685–19696 (October 2023) 2

[14] Sun, Y., Liu, Y., Liu, X., Li, Y., Chu, W.S.: Rethinking domain generalization for face anti-spoofing: Separability and alignment (3 2023), URL http://arxiv.org/abs/2303.13662 2, 4

[15] Teed, Z., Deng, J.: RAFT: Recurrent All-Pairs Field Transforms for Optical Flow (Extended Abstract). IJCAI International Joint Conference on Artificial Intelligence pp. 4839–4843 (2021), ISSN 10450823, https://doi.org/10.24963/ijcai.2021/662 8, 13

[16] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Val Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: The European Conference on Computer Vision (ECCV) (2016) 5

[17] Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., Wang, Z.: Domain generalization via shuffled style assembly for face anti-spoofing (3 2022), URL http://arxiv.org/abs/2203.05340 2, 5

[18] Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., Zhou, F., Lei, Z.: Deep spatial gradient and temporal depth learning for face anti-spoofing. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 5041–5050 (2020), ISSN 10636919, https://doi.org/10.1109/CVPR42600.2020.00509 2, 3, 13

[19] Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security **10**, 746–761 (4 2015), ISSN 15566013, https://doi.org/10.1109/TIFS.2015.2400395 5

[20] Xu, X., Xiong, Y., Xia, W.: On Improving Temporal Consistency for Online Face Liveness Detection (jun 2020), URL http://arxiv.org/abs/2006.06756 3, 4, 5, 6, 10

[21] Xu, Z., Li, S., Deng, W.: Learning temporal features using LSTM-CNN architecture for face anti-spoofing. Proceedings - 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015 pp. 141–145 (2016), https://doi.org/10.1109/ACPR.2015.7486482 2, 3, 5

[22] Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing (8 2014), URL http://arxiv.org/abs/1408.5601 2

[23] Yang, S., Wang, W., Xu, C., He, Z., Peng, B., Dong, J.: Exposing Fine-Grained Adversarial Vulnerability of Face Anti-Spoofing Models. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, vol. 2023-June, pp. 1001–1010 (2023), ISBN 9798350302493, ISSN 21607516, https://doi.org/10.1109/CVPRW59228.2023.00107, URL https://github.com/Songlin1998/SpoofGAN 2

[24] Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016), ISSN 10709908, https://doi.org/10.1109/LSP.2016.2603342 5

[25] Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. pp. 26–31 (2012), ISBN 9781467303941, https://doi.org/10.1109/ICB.2012.6199754 5

[26] Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Sheng, K., Ding, S., Ma, L.: Generative domain adaptation for face anti-spoofing (7 2022), https://doi.org/10.1007/978-3-031-20065-6_20, URL http://arxiv.org/abs/2207.10015http://dx.doi.org/10.1007/978-3-031-20065-6_20 2