Adaptive Transformers for Robust Few-shot Cross-domain Face Anti-spoofing

Hsin-Ping Huang[1], Deqing Sun[2], Yaojie Liu[2], Wen-Sheng Chu[2], Taihong Xiao[1], Jinwei Yuan[2], Hartwig Adam[2], and Ming-Hsuan Yang[1,2,3]

[1]University of California, Merced [2]Google Research [3]Yonsei University

We first discuss implementation and experiment details. Next, we present more ablation studies of our approach. Third, we provide additional visualization to validate the proposed design. Finally, we include additional discussions.

1 Implementation Details

Frame sampling strategy for Protocol 1. Following SSDG [6], we train the model using only one frame selected in each video, and evaluate the model using two frames in each video. Both the training and test set videos in the original dataset are considered as data in one domain, and there are no training and test splits for cross-domain evaluation. We hold out ten frames in each domain as the few-shot samples. We calculate the average probability for each video and the HTER, AUC and TPR@FPR=1% scores are obtained based on videos instead of frames. We use TPR@FPR=1% as the metric instead of TPR@FPR=0.1% since the test sets contain only hundreds of images.

Frame sampling strategy for Protocol 2. We train the model using ten frames equidistantly sampled from each video, and evaluate the model using ten frames in each video for WMCA and CeFA. We use all frames from the SURF dataset. We split the data into training and test sets for this protocol. As there are multiple spoof types in WMCA dataset, we use all videos for training, and use only print attacks, replay attacks and live videos for test. For CeFA and SURF datasets, we use all data as training set, and use the original validation set for test. We hold out ten frames in each domain as the few-shot samples. As SURF does not contain videos, we calculate the HTER, AUC and TPR@FPR=1% scores based on the probabilities of frames instead of videos. Based on the sampling strategy, the numbers of frames used in training/test are 16K/9K (WMCA), 60K/12K (CeFA) and 96K/9K (SURF).

Explanation about the benchmark. We follow the typical cross-domain evaluation setting that is widely used in face anti-spoofing literature [2,6,7,12–14,19–23,25]. We note that the zero-shot benchmark proposed in [15] has been previously retrieved and is no longer available. The benchmark proposed in [18] is less used in the FAS literature. The authors study a different topic of detecting novel live/spoof sub-categories, and their zero/few-shot settings use ten live and spoof samples from two different sub-categories for adaptation. The problem setup is different from our supervised few-shot domain adaptation task.

We report the comparison to [18] in Table 1 following their training/test protocols. Instead of using their meta-learning approach, we simply train the model with binary classification loss using the training set. At test time, we follow their zero/few-shot evaluation settings and use a balanced data batch formed by the few-shot samples and the samples from the training set to adapt the model. All models are initialized from the ImageNet-pretrained ViT and fine-tuned following the data splits in their benchmark. Our model performs favorably against [18], especially for the few-shot settings. No additional datasets are used in this experiment per ECCV guideline.

Table 1: **ACER (%) Comparison to [18]**. We report the results on the benchmarks proposed in [18]. Our model performs favorably against prior work, especially for the few-shot settings.

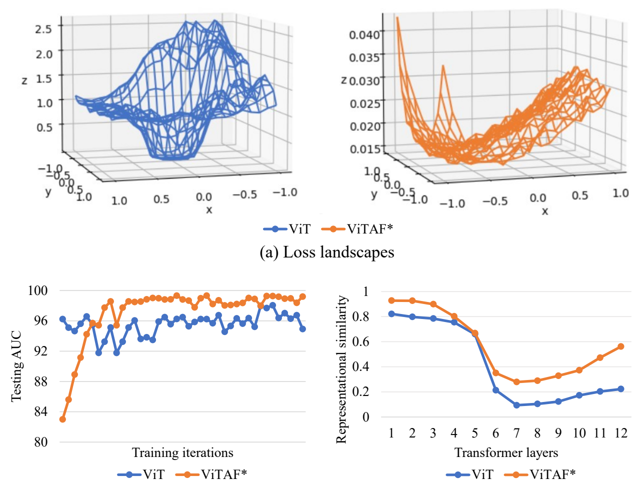
Benchmark	Method	0-shot	1-shot	5-shot
OULU-ZF	[18]	4.97±1.29	4.00±1.31	2.44±0.71
	Ours	**4.96±1.18**	**3.66±1.16**	**1.60±0.53**
SURF-ZF	[18]	30.97 ±1.28	28.75 ±1.49	27.27±1.25
	Ours	**28.53±2.48**	**22.50±2.33**	**21.60 ±2.95**

2 More Ablation Studies

2.1 Analysis of the Adaptive Module

As discussed in the paper, fine-tuning large models such as ViT with few samples usually causes instability. The proposed adaptive transformer model allows the transformer to operate on fewer parameters with a skip-connection to reduce optimization difficulty, thus achieving stability for mitigating domain gap in face anti-spoofing. Here we provide more analysis to justify the effectiveness of the proposed adaptive transformer model.

First, we plot the loss landscapes [11] of the naive ViT and our adaptive transformer in Fig. 1(a). The loss landscape by ViTAF* is wide and flat compared to the naive ViT, which indicates better generalization. Second, as discussed in Section 4.3 and Fig. 3 of the paper, the fine-tuning test performance of ViT model may fluctuate among different checkpoints even when the training loss converges. In Fig. 1(b), we show the learning curve of the ViT method and the proposed ViTAF* method. Although the best performance of ViT is already comparable to state-of-the-art, the performance fluctuates among different checkpoints and saturates since early iterations. In contrast, the test performance by the ViTAF* model gradually increases when it is trained for longer iterations, and there is no large fluctuation among the checkpoints. The learning curve validates the robust performance achieved by the ensemble adapters. Third, the skip-connection in the adapter module makes the output embedding mimic the input embedding, which yields representations with less deviation from the pre-trained ones and alleviates the instability caused by the catastrophic forgetting problem [4,9,17]. In Fig. 1(c), we plot the representational similarity [8] between the ImageNet-pretrained ViT and the fine-tuned ViT (blue line) versus the similarity between the ImageNet-pretrained ViT and the fine-tuned ViTAF* (orange line). The results show that the proposed model has higher feature similarity, thus indicating that ViTAF* better alleviates the catastrophic forgetting problem to achieve training stability.



(a) Loss landscapes

(b) Learning curve

(c) Representational similarity

Fig. 1: **Analysis of the adaptive module.** (a) The loss landscape of ViTAF* is wide and flat compared to naive ViT. (b) The test AUC of ViTAF* stably increases during the training process. (c) The representational similarity of ViTAF* is higher, indicating ViTAF* better alleviates the catastrophic forgetting problem.

2.2 Ablation Study of Number of Ensemble Adapters

In the experiments presented in the manuscript, we set the number of adapters in the ensemble adapter module as $K = 2$. In Table 2, we present the results of using different numbers of adapters in the ensemble module. While the ensemble module can be extended to include more than two adapters, we do not observe consistent improvements when increasing the number of adapters to $K = 3$ and $K = 4$. Thus, we keep $K = 2$ in the paper.

In addition, we include the results of ablating the cosine similarity loss L_{cos} in our framework. We observe that no matter how many adapters are used in the ensemble modules, the performance drop when we remove L_{cos}, which suggests that L_{cos} is essential for the multiple adapters to learn diverse feature and improve the accuracy.

2.3 Comparison to Feature-level Augmentation Methods

In our adaptive transformer model, we utilize the feature-wise transformation layer as we find its usefulness for cross-domain anti-spoofing task. We do not

Table 2: **Ablation study of number of ensemble adapters.** Increasing the number of adapters to $K = 3$ or $K = 4$ does not bring consistent improvements to the model. Removing the L_{cos} causes the AUC drop for all cases.

	OCI → M			**OMI → C**			**OCM → I**			**ICM → O**		
	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%
K=1	3.42	99.30	88.33	1.40	99.85	95.71	3.74	99.34	85.38	7.17	98.26	71.97
K=2 (Ours)	2.92	**99.62**	**91.66**	**1.40**	**99.92**	**98.57**	1.64	99.64	91.53	5.39	98.67	76.05
K=3	2.92	99.80	93.33	2.91	99.64	92.14	**1.35**	**99.88**	**98.46**	6.04	98.78	**83.38**
K=4	**1.58**	99.51	91.66	2.09	99.81	95.00	1.42	99.88	96.15	**5.36**	**98.87**	78.87
K=2 *w/o* L_{cos}	5.00	98.58	88.33	3.02	99.45	87.86	3.81	99.20	86.15	7.04	98.02	71.27
K=3 *w/o* L_{cos}	5.00	98.80	83.33	2.90	99.55	86.42	2.29	99.44	84.61	5.56	98.55	69.43
K=4 *w/o* L_{cos}	3.42	98.89	86.67	1.51	99.73	92.14	2.09	99.64	90.00	6.93	98.38	78.59

Table 3: **Comparison to feature-level augmentation methods.** We compare our model to representative feature-level augmentation approaches [5, 28]. The results show that our model is more effective for few-shot cross-domain anti-spoofing.

	OCI → M		**OMI → C**		**OCM → I**		**ICM → O**	
Method	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
SSDG	8.42	97.39	12.91	93.59	4.48	99.14	7.81	97.46
SSDG+DropBlock	3.41	99.13	14.30	93.79	6.12	98.75	9.56	97.05
SSDG+MixStyle	5.00	97.63	6.51	97.35	3.06	99.54	10.26	96.23
ViTAF*	**2.92**	**99.62**	**1.40**	**99.92**	**1.64**	**99.64**	**5.39**	**98.67**

claim that we proposed the most effective augmentation approach. In Table 3, we compare our methods to two representative feature-level augmentation methods for domain generalization, DropBlock [5] and MixStyle [28]. The two methods are not directly applicable to ViT and we add these approaches to our baseline SSDG for comparison. Although these feature augmentation methods [5, 28] achieve good performance on standard cross-domain object classification benchmark [10], the results show that our model is more effective for the face anti-spoofing domain.

2.4 Ablation Study of Increasing Few-shot Samples

In Fig. 2, we conduct an ablation study of increasing the number of few-shot target domain samples provided at training time. The x-axis shows the number of few-shot samples including 5-shot, 10-shot, 20-shot, 50-shot, 100-shot and all-shot. The y-axis shows the TPR@FPR=1% metric. Note that to conduct this experiment, we split the datasets into training and test sets, and thus the result is not directly comparable to the results in the main paper. We observe that the TPR@FPR=1% scores increase when more few-shot samples are included in training for all datasets. In addition, the TPR@FPR=1% score achieves 90% (which is almost saturated) when there are 5 shots (OCI → M), 10 shots (OMI → C), 20 shots (OCM → I), 100 shots (ICM → O), 50 shots (CS → W), 50 shots

(SW → C) and 20 shots (CW → S). The results suggest that including less than 5% samples in the target domain at training time can achieve good performance.

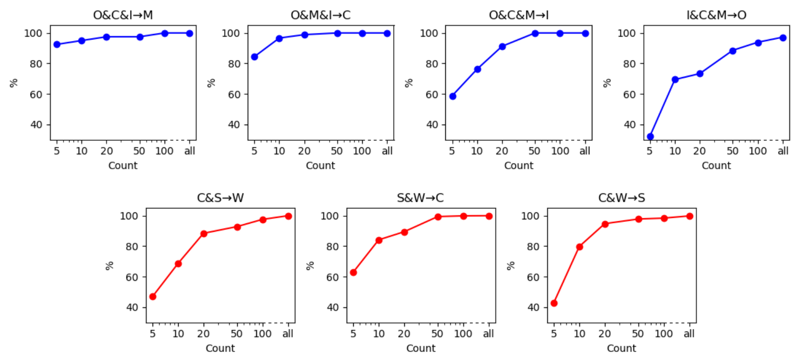


Fig. 2: **Ablation study of increasing few-shot samples.** We conduct ablation study of increasing the number of few-shot target domain samples provided at training time. The x-axis shows the number of few-shot samples from 5-shot to 100-shot and all-shot in log scale. The y-axis shows the TPR@FPR=1% metric.

2.5 Intra-database Evaluation

To continue the discussion in the previous section, we provide the intra-database results on four protocols of the Oulu-NPU dataset in Table 4. Although our approach does not aim at intra-database evaluation where the training/test sets are from the same domain, our method achieves performance comparable to the state-of-the-art method DC-CDN [24] for protocol 2 and 3, and it obtains favorable results for protocol 1 and 4. No additional datasets are used in this experiment per ECCV guideline.

Table 4: **Intra-database evaluation on Oulu-NPU dataset.** Our method is comparable to the state-of-the-art method for protocol 2 and 3, and it achieves favorable results for protocol 1 and 4.

Prot.	Method	APCER(%)↓	BPCER(%)↓	ACER(%)↓
	CDCN [26]	0.4	1.7	1.0
1	DC-CDN [24]	0.5	0.3	**0.4**
	Ours	0.4	1.8	1.1
	CDCN [26]	1.5	1.4	1.5
2	DC-CDN [24]	0.7	1.9	1.3
	Ours	1.1	1.1	**1.1**
	CDCN [26]	2.4	2.2	2.3
3	DC-CDN [24]	2.2	1.6	**1.9**
	Ours	0.4	3.3	**1.9**
	CDCN [26]	4.6	9.2	6.9
4	DC-CDN [24]	5.4	2.5	**4.0**
	Ours	0.0	15.0	7.5

2.6 Ablation Study of Using Different 5-shot Samples

In our experiments in the paper, the 5-shot samples of the target domains are selected randomly and fixed for all the experiments. In Table 5, we provide an ablation study of using different sets of 5-shot samples randomly selected from the target domain. We conduct the experiments for five runs and report the average performance and standard deviation.

The standard deviation of AUC score in five runs are: 0.27 (OCI $\rightarrow$ M), 0.50 (OMI $\rightarrow$ C), 0.98 (OCM $\rightarrow$ I), 0.36 (ICM $\rightarrow$ O), 0.33 (CS $\rightarrow$ W), 0.64 (SW $\rightarrow$ C), 1.87 (CW $\rightarrow$ S). Although the performance varies due to the selection of 5-shot samples, the variation is not large for most datasets. There are higher variations for Idiap (I), CeFA (C) and SURF (S). The datasets CeFA (C) and SURF (S) have 10× more examples than other datasets and the 5-shot samples cover only 0.1% of the data. Therefore, using different sets of 5-shot samples leads to a larger performance variation for SW $\rightarrow$ C and CW $\rightarrow$ S. On the other hand, OCM $\rightarrow$ I also has a larger performance variation caused by the selection of 5-shot samples. This might be due to the low resolution (480P) of the Idiap (I) dataset which causes some live videos to have even worse visual quality than spoof videos, and the recording environments are less diverse than other datasets. Therefore, if the selected 5-shot samples are outliers that do not cover the common live/spoof types, the model will be biased and the performance degrades. We note that although in Protocol 2 there are seven kinds of spoof attacks in the WMCA (W) dataset, we have only used replay and print attacks for testing. Thus, the 5-shot selection strategy is not an issue for CS $\rightarrow$ W as shown in Table 5. More details about the frame sampling strategy are in Section 1. It is interesting to study which samples obtained from the new domain are most beneficial for the few-shot adaptation, which we leave as future work.

Table 5: **Ablation study of using different 5-shot samples.** In our experiments in the paper, the 5-shot samples of each dataset are selected randomly and fixed for all the experiments. Here we provide an ablation study of using different sets of 5-shot samples randomly selected in the target domain. We conduct the experiments for five runs and report the average results and standard deviation.

	OCI → M			**OMI → C**			**OCM → I**			**ICM → O**		
	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%
Mean	2.37	99.55	91.53	3.14	99.34	86.64	5.60	98.68	72.77	8.33	97.68	64.79
Std	1.26	0.27	3.69	1.47	0.50	9.00	3.15	0.98	17.08	0.74	0.36	5.42

	CS → W			**SW → C**			**CW → S**		
	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%
Mean	4.31	99.32	86.51	6.51	98.20	68.76	9.31	96.15	66.76
Std	1.19	0.33	6.58	1.50	0.64	10.38	2.65	1.87	12.84

2.7 Ablation Study of Excluding CelebA-Spoof

Table 6 shows the ablation study of excluding CelebA-Spoof [27] from the source datasets. We present the results of our 0-shot ViT and 5-shot ViTAF model. ViT† and ViTAF† denote the model trained without using CelebA-Spoof in the source datasets. Excluding CelebA-Spoof causes the performance drop in all target datasets for the 0-shot ViT model, and six out of seven target datasets for the 5-shot ViTAF model. These results show that including CelebA-Spoof in the source datasets increases the diversity of training data and helps learn a better representation.

The 0-shot ViT† has an average 6.38 AUC performance drop compared to 0-shot ViT, while 5-shot ViTAF† has a 1.31 AUC drop compared to 5-shot ViTAF. The results suggest that when there are no target domain samples provided at training time, the diversity of source datasets highly affects the results. On the other hand, when there are few-shot samples, excluding CelebA-Spoof does not cause a drastic performance drop, though still a moderate difference, which shows the effectiveness of CelebA-Spoof dataset on improving the generalizability of the model.

Table 6: **Ablation study of excluding CelebA-Spoof.** We present the ablation study of excluding CelebA-Spoof [27] from the source datasets. We present the results of our 0-shot ViT and 5-shot ViTAF model. ViT† and ViTAF† denote the model trained without using CelebA-Spoof as the supplementary source dataset.

	Method	**OCI→M**			**OMI→C**			**OCM→I**			**ICM→O**		
		HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%
0-shot	ViT	1.58	99.68	96.67	5.70	98.91	88.57	9.25	97.15	51.54	7.47	98.42	69.30
	ViT†	4.75	98.79	68.33	15.70	92.76	36.43	17.68	86.66	50.77	16.46	90.37	24.23
5-shot	ViTAF	3.42	99.30	88.33	1.40	99.85	95.71	3.74	99.34	85.38	7.17	98.26	71.97
	ViTAF†	4.75	98.59	80.00	4.19	98.59	57.86	3.28	99.27	76.92	10.74	95.70	51.13

	Method	**CS→W**			**SW→C**			**CW→S**		
		HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%
0-shot	ViT	7.98	97.97	73.61	11.13	95.46	47.59	13.35	94.13	49.97
	ViT†	21.04	89.12	30.09	17.12	89.05	22.71	17.16	90.25	30.23
5-shot	ViTAF	4.51	99.44	88.23	7.21	97.69	70.87	11.74	94.13	50.87
	ViTAF†	4.91	98.78	75.95	13.56	93.68	30.90	12.63	94.21	55.03

2.8 Ablation Study of Alternative Transfer Learning Strategies

As discussed in Section 3.2 in the paper, one straightforward transfer learning strategy is to train a classifier on top of features extracted by the ViT backbone pre-trained on ImageNet [3] using anti-spoofing data. Another common strategy is to freeze a majority of the backbone and partially fine-tune the network. In Table 7, we investigate the transfer learning strategies on ViT models by comparing several alternatives: (1) ViT (fixed): fixing the ViT backbone and fine-tuning only the MLP head; (2) ViT (fine-tune last four): fine-tuning the last four layers along with the MLP head; (3) ViT (fine-tune last eight): fine-tuning the last eight layers.

The result shows that our transfer learning strategy outperforms the other alternatives where only parts of the network are fine-tuned. It also validates that fine-tuning the introduced ensemble adapters and feature-wise transformation layers effectively adapts the features of ViT to the anti-spoofing tasks. In addition, only fine-tuning the MLP head on top of a fixed ViT backbone leads to degraded performance, suggesting that ImageNet-pretrained ViT features are high-level thus cannot be directly used for anti-spoofing tasks where the subtle low-level information is crucial.

Table 7: **Ablation study of alternative transfer learning strategies.** The alternative strategies include: (1) ViT (fixed): fixing the ViT backbone and fine-tuning only the MLP head; (2) ViT (fine-tune last four): fine-tuning the last four layers along with the MLP head; (3) ViT (fine-tune last eight): fine-tuning the last eight layers.

	OCI → M			**OMI → C**			**OCM → I**			**ICM → O**		
	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%
ViT (fixed)	13.66	93.55	45.00	21.39	87.23	38.57	20.81	86.89	30.77	19.44	88.61	26.06
ViT (fine-tune last four)	6.83	98.00	68.33	10.69	95.52	49.29	9.32	95.99	52.31	11.69	94.86	50.99
ViT (fine-tune last eight)	3.41	98.63	80.00	5.69	98.33	70.71	6.12	98.31	66.15	12.03	95.50	40.56
ViTAF*	**2.92**	**99.62**	**91.66**	**1.40**	**99.92**	**98.57**	**1.64**	**99.64**	**91.53**	**5.39**	**98.67**	**76.05**

2.9 ViTF in 0-shot Setting

In the proposed framework, both the ensemble adaptor modules and the FWT layers are specifically designed and suitable for the few-shot domain adaptation problem. On the other hand, the ViTF model with FWT layers can be served as a general feature-wise augmentation method and applied to the 0-shot setting when there are no target domain samples available. We include the results of ViTF model in 0-shot setting in Table 8. We observe that the ViTF model does not have improvements compared to ViT for the setting **ICM→O**. A possible reason is **O** has a more diverse distribution than other datasets in protocol 1. Thus, the data augmentation on the source datasets **ICM** does not bring improvement to the target **O**.

Table 8: **ViTF in 0-shot setting.** Including the FWT layer does not bring improvement to ICM→O in 0-shot setting.

	OCI → M		**OMI → C**		**OCM → I**		**ICM → O**	
Method	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
SSDG	6.58	97.21	12.91	93.92	7.01	98.28	12.47	94.87
ViT	**1.58**	99.68	5.70	98.91	9.25	97.15	**7.47**	**98.42**
ViTF	2.20	**99.74**	**4.89**	**99.21**	**6.12**	**98.55**	7.89	97.58

2.10 Runtime Analysis and Network Comparison

In Table 9, we show the training and inference time of the proposed method, the details of network sizes and the computational complexity. The analysis is conducted on a desktop machine equipped with an Nvidia 2080Ti GPU. We show that the network parameters and the computational complexity of the ViTAF* model increase only ≈ 5.5% compared to the naive ViT model. The analysis shows that the proposed ensemble adapter modules and the feature-wise transformation layers are lightweight. In practice, the training time and inference time of ViTAF* increase only 15% compared to the ViT model. Overall, it takes three hours to finish the training/evaluation for protocol 1, and six hours for protocol 2.

Table 9: **Run-time analysis.** We present the training, inference time, network sizes and computational complexity of the proposed method.

Pre-training time (seconds / per iteration)	ViT/ViTAF*	0.17
Fine-tuning time (seconds / per iteration)	ViT	0.18
	ViTAF*	0.21 (+15%)
Inference time (seconds / per input frame)	ViT	0.013
	ViTAF*	0.015 (+15%)
Number of parameters	ViT	86.39M
	ViTAF*	91.15M (+5.5%)
FLOPs	ViT	33.69G
	ViTAF*	35.56G (+5.5%)

3 Visualization

ROC curves. The ROC curves represent equivalent information as the AUC scores provided in the paper. In Fig. 3 we plot the ROC curves of our method and the baseline model SSDG for reference.

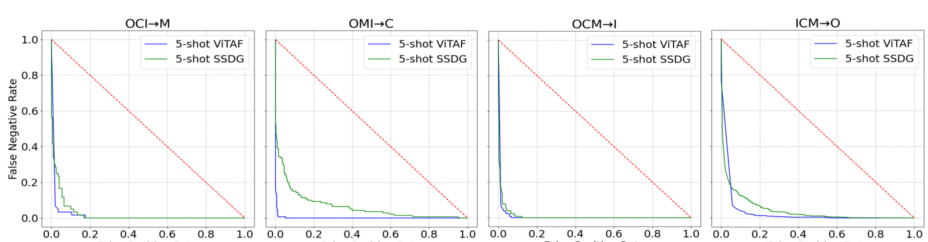


Fig. 3: **ROC curves.** We compare the ROC curves of our method and the SSDG baseline for protocol 1.

Failure cases. In Fig. 4, we provide additional failure case analysis for Protocol 2. The live faces misclassified as spoof faces are shown in blue boxes and the spoof faces misclassified as live faces are shown in red boxes. The live faces misclassified as spoof faces (blue) are in dark light conditions (top left), or have bad visual quality and larger pose changes (top right). In addition, the live faces with darker skin are more challenging (example shown in the top middle). The spoof faces misclassified as live faces (red) are mostly replay attacks without obvious spoof cues for **CS** $\rightarrow$ **W**. As for **SW** $\rightarrow$ **C**, faces printed on clothes in either indoor or outdoor light conditions are the most challenging spoof types as there are no obvious spoof cues. The most challenging spoof type for **CW** $\rightarrow$ **S** is the person holding curved photos with good visual quality. Since images in the SURF dataset generally have low resolution and bad visual quality, the spoof faces with better quality are easily misclassified as live.

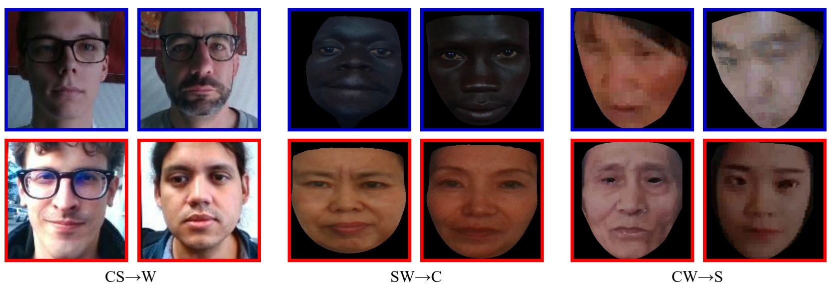


Fig. 4: **Failure case analysis for Protocol 2.** We provide additional failure case analysis. The live faces misclassified as spoof faces are shown in blue boxes and the spoof faces misclassified as live faces are shown in red boxes.

Attention maps. As shown in Fig. 5, we visualize the attention maps of different transformer models on spoof images using Transformer Explainability [1]. We observe that different regions are highlighted by transformers to make predictions for different spoof face domains. For example, transformers make predictions mainly based on the paper boundaries or reflection on screens for paper and replay attack in **CS** $\rightarrow$ **W**. For attacks printed on clothes, transformers focus on the wrinkles in **SW** $\rightarrow$ **C**. As for the paper attacks where eye or nose regions are cut out in **CW** $\rightarrow$ **S**, transformers pay more attention to the cut regions (holes) on the spoof medium. Moreover, our methods can better capture the spoof cues compared to the naive ViTs, as the attention region given by our models is more conspicuous.

In Fig. 6, we provide additional visualization of attention. Live or spoof faces are shown in blue or red boxes. Transformer focuses on reflection light on noses for live faces in **OCI** $\rightarrow$ **M** and **OMI** $\rightarrow$ **C**. The reflection light on noses is usually a cue of live faces. In addition, transformer focuses on reflections on glasses and boundaries of the painting on the background for **OCM** $\rightarrow$ **I**. Though glasses or painting on the background are not spoof cues, the reflection on glasses is

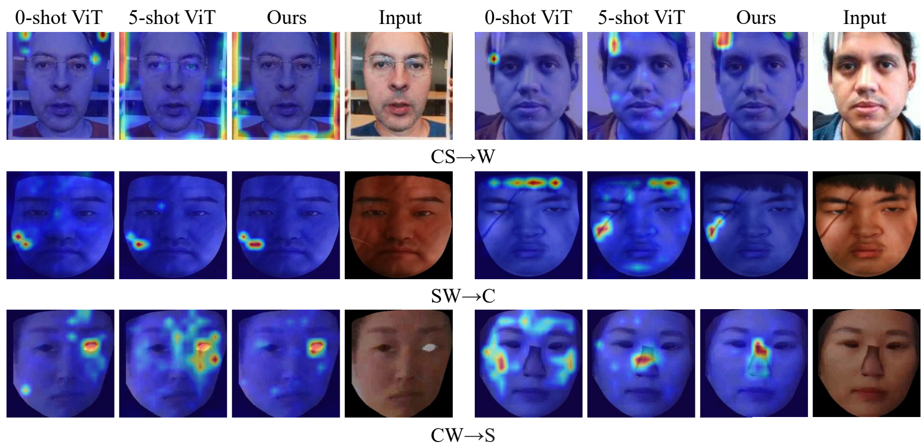


Fig. 5: **Transformer attention on spoof images for Protocol 2.** We visualize the attention maps of transformers using Transformer Explainability [1] and make a comparison of naive ViT and our models. Transformers focus on paper boundaries or reflection on screens (**CS** → **W**), clothes wrinkles (**SW** → **C**), holes at eye or nose regions (**CW** → **S**). Our models generate more accurate and conspicuous attention maps to capture spoof cues compared with others.

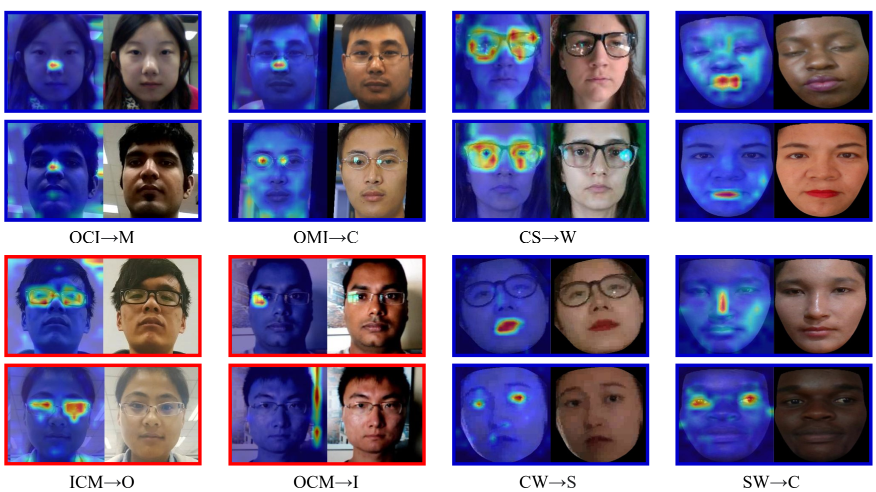


Fig. 6: **Additional results.** We provide additional visualization of the attention maps. Live or spoof faces are shown in blue or red boxes.

similar to the reflections caused by replay attacks, and the painting boundaries look similar to the boundaries of spoof mediums. Our model focuses on glasses region for **OMI** $\rightarrow$ **C**, **ICM** $\rightarrow$ **O** and **CS** $\rightarrow$ **W** as well. As for the live faces in **CW** $\rightarrow$ **S** and **SW** $\rightarrow$ **C**, transformer model mainly focuses on eyes, nose, and mouth region which usually show cues for live faces.

Feature visualization of Protocol 2. We present the t-SNE plot [16] of Protocol 2 in Fig. 7. Each plot indicates the features extracted from the model of each setting shown on the top. Each color in the plot shows the live/spoof samples in each dataset. We observe that features of WMCA (purple, brown) extracted from model **CS** $\rightarrow$ **W** (left) are well-separated, demonstrating that the model generalizes well to the target domain. In addition, features of CeFA (blue, orange) extracted from model **SW** $\rightarrow$ **C** (middle) are almost but not entirely separated, indicating there is still room of improvement for this model. On the other hand, features of SURF (green, red) extracted from model **CW** $\rightarrow$ **S** (right) are mixed. Since images in SURF have very low visual quality, the model trained on CW does not generalize well to S.

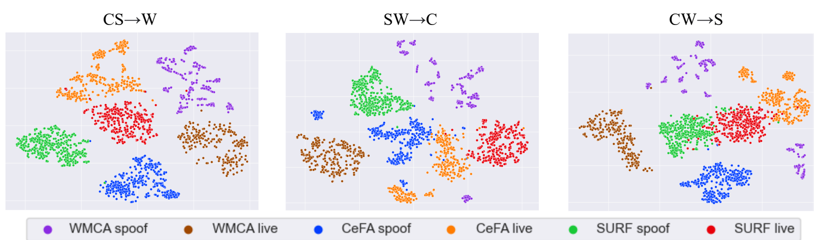


Fig. 7: **Feature visualization of Protocol 2.** We present the t-SNE plot of Protocol 2. Each plot indicates the features extracted from the model of each setting shown on the top. Each color in the plot shows the live/spoof samples in each dataset.

4 Discussions

Although adapter methods are recently developed in natural language processing (NLP), our paper aims at the face anti-spoofing problem in computer vision, which is significantly different from NLP. We develop the first method to successfully apply the adaptive module to vision transformers for face anti-spoofing domain, especially for the cross-domain FAS task. Our adaptive transformer model effectively adapts the pre-trained model to novel domains with only a few samples available. Therefore, it is a promising direction for face anti-spoofing domain where the data is usually hard to obtain.

It is challenging to apply the adaptive module to cross-domain face anti-spoofing task. We introduce the ensemble adaptive module with the cosine loss which is essential to achieve the stable and good performance. We believe that our findings are worth sharing with the face anti-spoofing community.

References

1. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
2. Chen, Z., Yao, T., Sheng, K., Ding, S., Tai, Y., Li, J., Huang, F., Jin, X.: Generalizable representation learning for mixture domain face anti-spoofing. In: Association for the Advancement of Artificial Intelligence (AAAI) (2021)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
4. Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., Smith, N.A.: Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305 (2020)
5. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. In: Neural Information Processing Systems (NeurIPS) (2018)
6. Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
7. Kim, T., Kim, Y.: Suppressing spoof-irrelevant factors for domain-agnostic face anti-spoofing. IEEE Access (2021)
8. Laakso, A., Cottrell, G.: Content and cluster analysis: Assessing representational similarity in neural systems. Philosophical Psychology **13**, 47 – 76 (2000)
9. Lee, C., Cho, K., Kang, W.: Mixout: Effective regularization to finetune large-scale pretrained language models. In: International Conference on Learning Representations (ICLR) (2020)
10. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: IEEE International Conference on Computer Vision (ICCV) (2017)
11. Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T.: Visualizing the loss landscape of neural nets. In: Neural Information Processing Systems (NeurIPS) (2018)
12. Liu, S., Lu, S., Xu, H., Yang, J., Ding, S., Ma, L.: Feature generation and hypothesis verification for reliable face anti-spoofing. In: Association for the Advancement of Artificial Intelligence (AAAI) (2022)
13. Liu, S., Zhang, K.Y., Yao, T., Bi, M., Ding, S., Li, J., Huang, F., Ma, L.: Adaptive normalized representation learning for generalizable face anti-spoofing. In: ACM International Conference on Multimedia (ACM MM) (2021)
14. Liu, S., Zhang, K.Y., Yao, T., Sheng, K., Ding, S., Tai, Y., Li, J., Xie, Y., Ma, L.: Dual reweighting domain generalization for face presentation attack detection. In: International Joint Conference on Artificial Intelligence (IJCAI) (2021)
15. Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X.: Deep tree learning for zero-shot face anti-spoofing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
16. van der Maaten, L., Hinton, G.: Viualizing data using t-sne. Journal of Machine Learning Research (JMLR) **9**, 2579–2605 (2008)
17. Phang, J., Févry, T., Bowman, S.R.: Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. arXiv preprint arXiv:1811.01088 (2018)
18. Qin, Y., Zhao, C., Zhu, X., Wang, Z., Yu, Z., Fu, T., Zhou, F., Shi, J., Lei, Z.: Learning meta model for zero-and few-shot face antispoofing. In: Association for the Advancement of Artificial Intelligence (AAAI) (2020)

19. Saha, S., Xu, W., Kanakis, M., Georgoulis, S., Chen, Y., Paudel, D.P., Van Gool, L.: Domain agnostic feature learning for image and video based face anti-spoofing. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2020)
20. Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
21. Shao, R., Lan, X., Yuen, P.C.: Regularized fine-grained meta face anti-spoofing. In: Association for the Advancement of Artificial Intelligence (AAAI) (2020)
22. Wang, G., Han, H., Shan, S., Chen, X.: Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
23. Wang, J., Zhang, J., Bian, Y., Cai, Y., Wang, C., Pu, S.: Self-domain adaptation for face anti-spoofing. In: Association for the Advancement of Artificial Intelligence (AAAI) (2021)
24. Yu, Z., Qin, Y., Zhao, H., Li, X., Zhao, G.: Dual-cross central difference network for face anti-spoofing. In: International Joint Conference on Artificial Intelligence (IJCAI) (2021)
25. Yu, Z., Wan, J., Qin, Y., Li, X., Li, S., Zhao, G.: Nas-fas: Static-dynamic central difference network search for face anti-spoofing. IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI) **43**, 3005–3023 (2021)
26. Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
27. Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., Liu, Z.: Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In: European Conference on Computer Vision (ECCV) (2020)
28. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: International Conference on Learning Representations (ICLR) (2021)