Domain Generalization via Shuffled Style Assembly for Face Anti-Spoofing Supplementary Material

Zhuo Wang[1] Zezheng Wang[2*] Zitong Yu[3] Weihong Deng[1*]
Jiahong Li[2] Tingting Gao[2] Zhongyuan Wang[2]

[1]Beijing University of Posts and Telecommunications [2]Kuaishou Technology [3]CMVS, University of Oulu

{wz2019, whdeng}@bupt.edu.cn zitong.yu@oulu.fi

{wangzezheng, lijiahong, wangzhongyuan}@kuaishou.com tinagao2019@gmail.com

1. Overview

More network architecture details, experimental results, further analysis of proposed benchmarks, discussion and potential future work are provided in this supplementary material. Specifically, Section 2 provides network architecture details of SSAN-M and SSAN-R. Section 3 provides the results of the cross-type testing on CASIA-MFSD [24], Replay-Attack [3], and MSU-MFSD [18]. Section 4 provides further descriptions and analyses of proposed benchmarks. Section 5 provides further discussion about our method. Section 6 describes the existing problems and future work direction.

2. Network Architecture

The detailed structures of SSAN-M and SSAN-R are shown in Fig. 1. For the structure, there exist differences in the feature generator and classifier between them. For the loss function, SSAN-M and SSAN-R have different forms of L_{cls} according to their supervision approaches. Specifically, depth supervision and L_{Depth} are usually used in DepthNet [11] based architectures SSAN-M, while binary supervision and L_{BCE} are used in ResNet-18 [5] based architectures SSAN-R. Their formulas are shown as follow:

$$L_{BCE} = \frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log x_i + (1-y_i)\cdot \log\left(1-x_i\right)\right], \tag{1}$$

where N is the number of the samples and y_i is the binary label of sample x_i.

$$L_{Depth} = \frac{1}{N}\sum_{i=1}^{N}\left\|D_P\left(i\right) - D_G\left(i\right)\right\|_2, \tag{2}$$

where N is the number of the samples, $D_P(i)$ and $D_G(i)$ represent the predicted depth map and ground-truth depth map, respectively. Thus, mean squared errors are calculated to measure the difference between them.

[*] denotes the corresponding author.

3. Cross-Type Testing

In the intra- and inter- dataset scenarios, different attack types can be considered as unique data fields, thus the performance of facing unknown attacks also reflects the cross-domain capability of the algorithms. Following the protocol proposed in [1], we use CASIA-MFSD [24], Replay-Attack [3], and MSU-MFSD [18] to perform intra-dataset cross-type testing and inter-dataset cross-type testing. All of them are small-scale datasets and contain a variety of common attack methods, including photo and video attacks.

Intra-Dataset Cross-Type Testing. As shown in Table 1, we adopt the Leave-One-Out (LOO) strategy for different attack types in the same dataset to evaluate the robustness of encountering unknown attacks. Five state-of-the-art methods are listed for comparison and our proposed methods achieve the best performance, which indicates the capability to process unknown presentation attacks.

Inter-Dataset Cross-Type Testing. The data distribution in the same dataset is similar. However, in reality, unknown presentation attacks usually appear in different domains. As shown in Table 1, we further estimate the performance of our method with inter-dataset protocols in [1]. Three methods are listed for comparison, including SVM1+IMQ [1], CDCN [21], and CDCN++ [21]. In this testing, our method retains the best performance compared with other methods, which demonstrates that our method has a great adaptation capability towards domain and unknown attacks.

4. Large-Scale FAS Benchmarks

4.1. Implementation Details

Datasets. Twelve datasets are used in the large-scale benchmark, including CASIA-SURF [22], WMCA

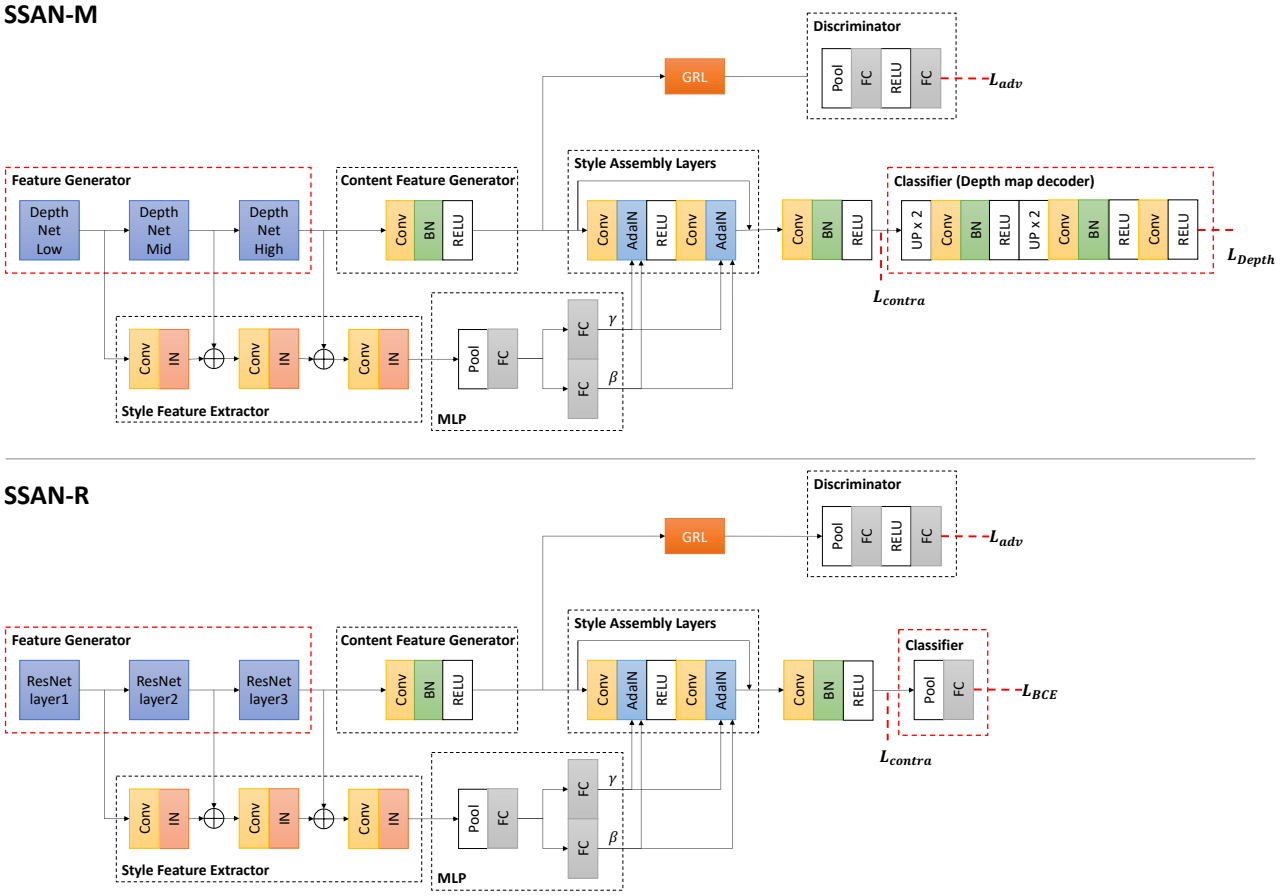

Figure 1. The detailed architectures of SSAN-M and SSAN-R. Specifically, embedding layers of DepthNet [11] are as the feature generator in SSAN-M while embedding layers of ResNet-18 [5] are as the feature generator in SSAN-R. Depth supervision is used in SSAN-M while binary supervision is used in SSAN-R to optimize the models. Thus, their corresponding L_{cls} are L_{Depth} and L_{BCE}, respectively. The red dashed line highlights their unique modules.

Table 1. AUC (%) of the intra-dataset cross-type and inter-dataset cross-type testing on CASIA-MFSD, Replay-Attack, and MSU-MFSD.

Method	Protocol	CASIA-MFSD			Replay-Attack			MSU-MFSD			Overall
		Video	Cut photo	Warpped Photo	Video	Digital Photo	Printed Photo	Printed Photo	HR Video	Mobile Video	
DTN [12]	Intra	90.00	97.30	97.50	99.90	99.90	99.60	81.60	99.90	97.50	95.90±6.20
CDCN [21]		98.48	99.90	99.80	100.00	99.43	99.92	70.82	100.00	99.99	96.48±9.64
CDCN++ [21]		98.07	99.90	99.60	99.98	99.89	99.98	72.29	100.00	99.98	96.63±9.15
BCN [19]		99.62	100.00	100.00	99.99	99.74	99.91	71.64	100.00	99.99	96.77±9.99
NAS-FAS [20]		99.62	100.00	100.00	99.99	99.89	99.98	74.62	100.00	99.98	97.12±8.94
SSAN-M (Ours)		97.65	99.52	98.68	100.00	100.00	99.83	86.88	100.00	99.25	**97.98±3.99**
SVM1+IMQ [1]	Inter	88.41	75.14	75.23	88.21	71.20	56.41	56.62	71.12	49.75	70.23±12.69
CDCN [21]		72.20	79.31	84.22	97.73	94.89	96.70	74.25	98.88	100.00	87.69±10.56
CDCN++ [21]		73.12	76.64	78.36	96.66	92.92	97.67	74.25	98.13	100.00	87.53±10.90
SSAN-M (Ours)		73.20	75.27	82.69	97.48	89.26	96.04	79.69	99.75	98.75	**88.01±9.93**

[4], HKBU-MARs V2 [10], CeFA [9], MSU-MFSD [18], OULU-NPU [2], CelebA-Spoof [23], CASIA-MFSD [24], REPLAY-ATTACK [3], WFFD [6], SiW [11], and Rose-Youtu [8]. For image data, we utilize all images of them. For video data, we extract frames of them at specific intervals to ensure similar data quantities. We firstly convert different raw data into image format, then merge them into a larger data distribution to simulate the realistic spectacles. Thus, the numbers of live / spoof images during training are 738624 / 1388138, 316227 / 666380, and 422397 / 721758 in protocol 1, 2 1, and 2 2, respectively. Their detailed information is shown in Table 2.

Protocols. For different testing scenarios, we set up corresponding testing protocols, and detailed information is

Table 2. Details of the datasets we use in the large-scale benchmark.

Dataset	Raw Format	Attack Types	Interval	Images Num	
				Train	Test
CASIA-SURF [22]	Image	Print	-	28876	56903
WMCA [4]	Video	Print, Replay, Mask	1	38293	664
HKBU-MARs V2 [10]	Video	Mask	1	254300	1328
CeFA [9]	Video	Print, Replay, Mask	1	387539	44112
MSU-MFSD [18]	Video	Print, Replay	1	33585	1280
OULU-NPU [2]	Video	Print, Replay	1	240014	14400
CelebA-Spoof [23]	Image	Print, Replay, Mask	-	456509	64884
CASIA-MFSD [24]	Video	Print, Replay	1	45085	2880
REPLAY-ATTACK [3]	Video	Print, Replay	1	92951	3840
WFFD [6]	Image	Waxworks	-	5280	1756
SiW [11]	Video	Print, Replay	3	274856	11920
Rose-Youtu [8]	Video	Print, Replay, Mask	3	269474	5560
Total	Image	-	-	2126762	209527

shown in Table 3. Significantly, for protocol 2, we divide these datasets into two piles according to their data quantities and attack types. Thus, protocols 2_1 and 2_2 are set to evaluate the performance of our methods between multiple datasets.

Metrics. In the real-world scenarios, there exist extensive live faces but few spoof faces, thus the Recall is usually used to evaluate the performance of the algorithms in reality. On the other hand, the metrics TPR@FPR at specific values have been widely used in face verification, such as IJB-C [13]. Thus, in the large-scale FAS benchmarks, we gather all live faces as negative cases while partial spoof faces as positive cases to calculate their TPR@FPR on each dataset. Then, the mean and variance of them are used for an overall evaluation.

4.2. Experimental Analysis

In the manuscript, we compare our method with different network structures ($i.e.$, CNN [5] and Transformer [16]) and some recent state-of-the-art methods ($i.e.$, CDCN [21] and SSDG [7]). In terms of the quantitative results, our method has achieved the best performance, compared with other methods. To make a further analysis, we draw ROC curves of each testing set on all protocols, as shown in Fig. 3, 4, and 5.

Fig. 3 describes the performance of different methods on protocol 1. It can be observed that our method outperforms the other methods in most datasets, which demonstrates the effectiveness of our method in the intra-dataset testing scenario. However, CDCN obtains the worst performance due to the limitation of model capability (parameter quantity), as shown in Table 5. It is worth noting that all models achieve nearly 100% accuracy in WMCA, which indicates there exist more universal cases in this dataset, thus great fitting can be obtained by it.

Fig. 4 describes the performance of different methods on protocol 2_1. It can be observed that our method shows competitive performance in most datasets, which demonstrates the effectiveness of our method in the cross-dataset testing scenario. However, SSDG-R obtains the worst performance in this setting, which can be attributed to the following reasons: 1) In large-scale data scenarios, attacks images from different domains may share some common distributions, thus the Asymmetric Triplet Mining proposed in [7] may confuse the optimization of different attack data from different domains; 2) There exist a broad distribution for live faces in the real-world scenarios where have extensive live faces but few spoof faces, as shown in Fig. 6, thus the optimization on the complete representations of single-side adversarial learning may be difficult in the real-world scenarios.

Fig. 5 describes the performance of different methods on protocol 2_2. It can be observed that our method shows the best performance in most datasets, compared with the other methods, which demonstrates the effectiveness of our method in the cross-dataset testing scenario. However, CDCN obtains the worst performance due to the limitation of model capability (parameter quantity), as shown in Table 5. Besides, CDCN is mainly designed for intra-dataset testing without domain adaptation (DA) or domain generalization (DG) techniques, thus may encounter degradation when facing unseen data. On the other hand, it is worth noting that our method obtains awful performance on HKBU, which may be attributed to the following reasons: 1) There exist limit mask attack images in the training set, thus poor performance is almost obtained by each network structure; 2) There exist diverse materials for mask attacks, such as hard resin, silicone, paper, plastic and so on. Different materials have unique texture information, which may cause chaos in contrastive learning for stylized features. To overcome the above problems, more mask attacks need to be collected as the training set in future works.

4.3. Ablation Study

To further verify the superiority of our SSAN as well as the contributions of each component, we form multiple incomplete models by controlling different variables and measure their performance on the large-scale FAS benchmark. Their results are shown in Table 4. It can be observed that our final model can achieve the best perfor-

Table 3. Details of the protocol implementation in the large-scale benchmark.

Dataset	Protocol 1 (Live / Spoof)		Protocol 2_1 (Live / Spoof)		Protocol 2_2 (Live / Spoof)	
	Train	Test	Train	Test	Train	Test
CASIA-SURF [22]	738624 / 1388138	58095 / 39745	/		422397 / 721758	31190 / 39745
WMCA [4]		58095 / 104				31190 / 104
HKBU-MARs V2 [10]		58095 / 656				31190 / 656
CeFA [9]		58095 / 34512				31190 / 34512
MSU-MFSD [18]		58095 / 2160				31190 / 960
OULU-NPU [2]		58095 / 11520				31190 / 11520
CelebA-Spoof [23]		58095 / 45057	316227 / 666380	26905 / 45057	/	
CASIA-MFSD [24]		58095 / 960		26905 / 2160		
REPLAY-ATTACK [3]		58095 / 3200		26905 / 3200		
WFFD [6]		58095 / 878		26905 / 878		
SiW [11]		58095 / 8480		26905 / 8480		
Rose-Youtu [8]		58095 / 4160		26905 / 4160		

mance, which proves the effectiveness of each component.

Table 4. The ablation study on the large-scale FAS benchmarks. Only the results (%) under TPR@FPR=10% on each protocol are reported there.

Method	Protocols		
	1	2_1	2_2
SSAN-R w/o L_{adv}	97.67±6.73	61.20±23.49	62.80±25.22
SSAN-R w/o L_{contra}	94.92±6.26	57.85±21.47	55.66±29.98
SSAN-R w/o stop-grad	98.19±4.81	63.27±17.32	56.67±28.95
SSAN-R w/ hard-sup	97.77±5.02	58.23±22.06	60.46±27.21
SSAN-R w/ SCL	97.45±4.60	57.67±21.74	55.76±31.07
SSAN-R (Ours)	**98.31±4.19**	**63.61±21.69**	**64.54±28.36**

4.4. Model Efficiency

In Table 5, we list the number of parameters and MACs to compare the model size and computation efficiency between different methods. It can be observed that our method not only has modest parameters (8.204M) and computation (2.235GMac), but also obtains excellent performance on the existing and proposed benchmarks, which proves the efficiency of our method.

Table 5. The comparison of parameter quantity and computational complexity.

Model	Backbone	Parameters	MACs
ResNet18 [5]	-	11.178M	2.375GMac
Deit-T [16]	-	5.477M	1.075GMac
CDCN [21]	-	2.245M	47.428GMac
SSDG-R [7]	ResNet18	12.758M	2.904GMac
SSAN-R	ResNet18	8.204M	2.235GMac

4.5. Visualization

Features Visualization. To future analyze the feature space learned by our SSAN method, we visualize the feature distribution under each sub-datasets in the large-scale benchmark protocol 1 using t-SNE [17], as shown in Fig. 6. It is worth noting that there contain all live faces and partial spoof faces in each sub-dataset testing, which is more similar to the real-world scenarios. Thus, it can be observed the following phenomenons: 1) The features of living faces can access to a broader distribution, compared to that of spoof faces, due to the imbalance that extensive live faces but few spoof faces in the testing setting and realistic spectacles; 2) Our method can separate spoof images from the live ones effectively, which proves the superiority of our method among multiple datasets.

5. Discussion

The Assembled Features for Classification. The reasons: 1) The proposed contrastive learning is implemented on the assembled features, which is important for style features extraction; 2) Content features contain important semantic cues, such as facial landmarks, which are complementary to style characteristics. Fig. 5 (b) in the manuscript shows that the distribution of style features is compact, which indicates they are usually similar for intra-categories. Nevertheless, when different style features are applied to the corresponding content ones for the assembled representations, it will further enhance their distribution difference between living and spoofing shown in Fig. 5 (c) in the manuscript; 3) Ablation study on O&C&I to M is also conducted to prove it quantitatively (Only style features: 19.58% HTER and 90.38% AUC; Assembled features: 10.42% HTER and 94.76% AUC).

Comparison with Triplet Mining. There exist the following differences between our SSAN and typical methods of triplet mining: 1) Our method decouples the representation into content and style features by utilizing their unique properties, then assembles various pairs of them to conduct contrastive learning for DG. This is a new perspective to eliminate domain bias; 2) Our method is more suitable for large-scale data scenarios. Because it is relatively difficult to conduct triplet mining on the complete representations directly when facing mass data. But for separated features, their common properties can be better induced and used with the increase of data; 3) Experimental results have proven the superiority of our method compared with the previous methods. For example, our method obtains great improvements on protocol 2_1 of proposed benchmarks by 10.17%, 22.29%, and 6.52%, respectively, compared with SSDG [7] which is a representative work of triplet mining.

Thus, our method is different from the previous methods.

Wrong Analysis. Fig. 2 shows that most errors are caused by the challenging appearance, such as low- or high-light conditions, color distortions, or image blurring. These adverse effects may mask the difference between living and spoofing.

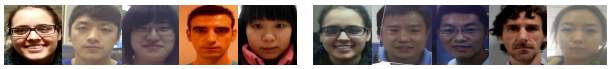

Figure 2. Examples of incorrect results. Left: Live; Right: Spoof.

6. Future Work

Though the preliminary experimental results have been obtained, there still exist some problems to be further studied, as follows:

Long-Tail Distribution. As shown in Table 2, there contain unbalanced quantities between different datasets, though specific sample intervals are used to alleviate this problem. This imbalance may lead to different performances in different datasets for intra-dataset testing shown in Fig. 3 and for cross-dataset testing shown in Fig. 4 and 5. Specifically, the long-tail data may be trapped in inferior performance under the large-scale benchmark because of unbalanced optimization. Therefore, the long-tail problems in FAS need to be further studied and developed.

Domain Partition. In the manuscript, adversarial learning is used to make generated content features indistinguishable for different domains. These domains mean different datasets, which may be suitable for the testing scenarios that contain several datasets such as OCIM [14, 15], but may suffer degradation when more datasets are provided, because there may exist overlap between different datasets. Our method is aimed to fetch close the content features of data in different domains, thus may suffer less impact from the above problems. However, methods that utilize the domain information for triplet supervision may be subjected to a serious degradation ($i.e.$, SSDG-R shown in Fig. 3, 4, and 5). Therefore, the soft approaches of domain partition by clustering methods need to be further explored in the testing scenarios containing multiple datasets.

Cross-Type Testing. In the large-scale benchmark, intra- and cross- dataset protocols have been proposed to evaluate the performance of the algorithms. However, the robustness of encountering unknown attacks is also important to be measured. Thus, the corresponding cross-type protocols will be designed for our large-scale benchmark in the future.

To sum up, the above problems constitute our future research directions.

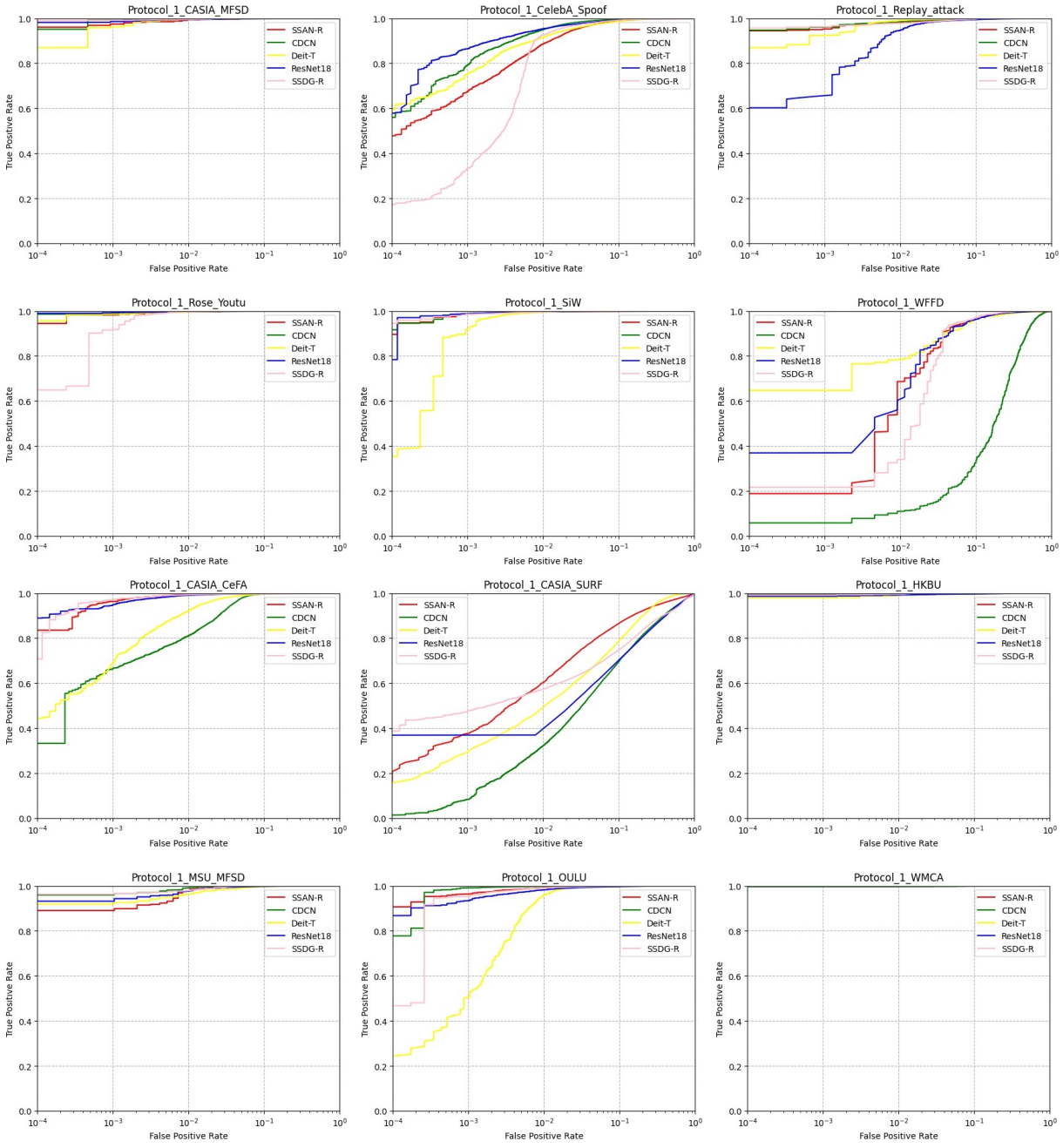

Figure 3. ROC curves of twelve testing sets for domain generalization on the large-scale FAS benchmark protocol 1.

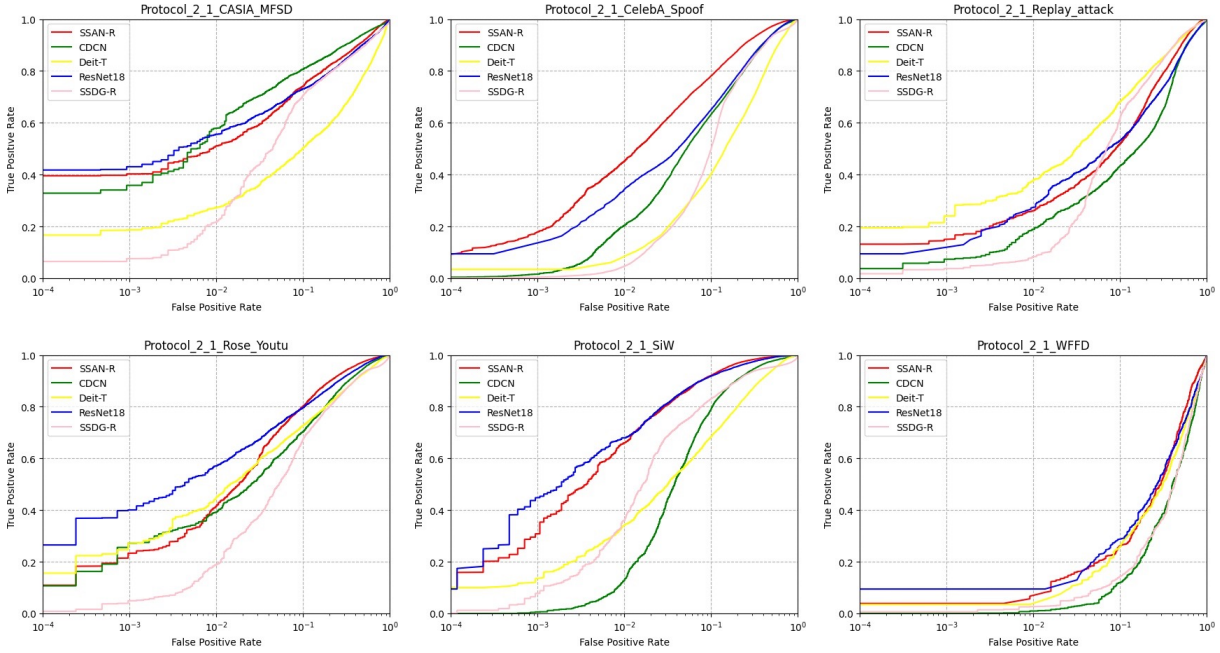

Figure 4. ROC curves of six testing sets for domain generalization on the large-scale FAS benchmark protocol 2_1.

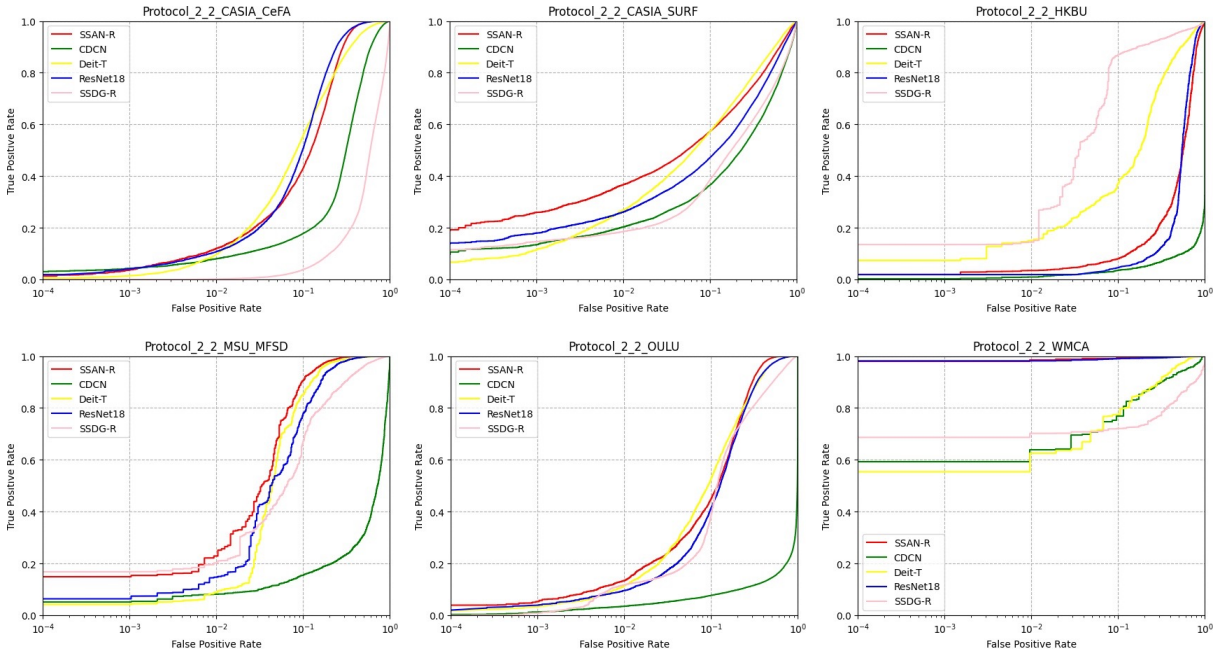

Figure 5. ROC curves of six testing sets for domain generalization on the large-scale FAS benchmark protocol 2_2.

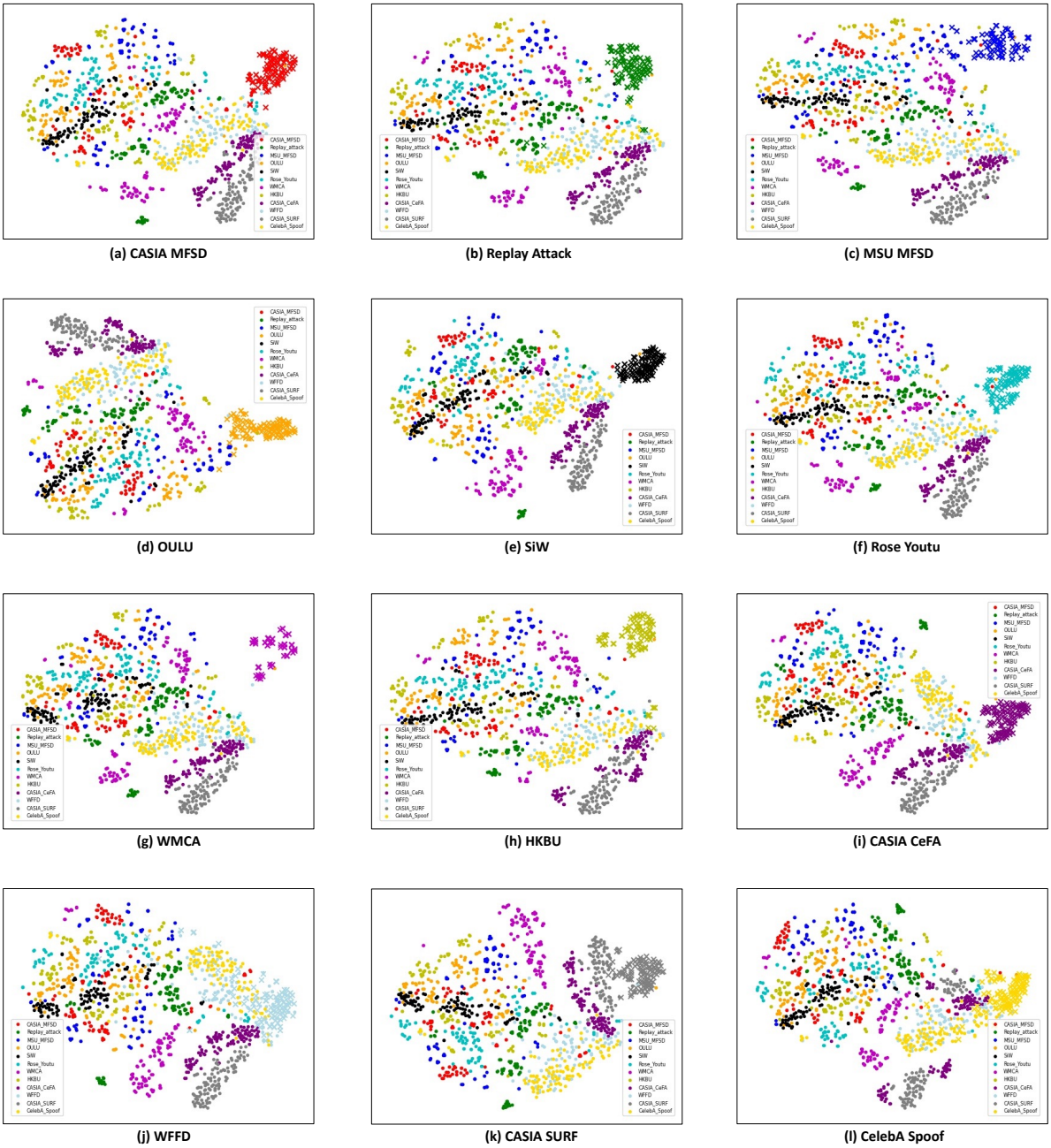

Figure 6. The t-SNE [17] visualizations of different features generated from SSAN-R under the large-scale benchmark protocol 1. In every sub-testings, all live faces are as negative cases while partial spoof faces in current datasets are as positive cases, thus TPR@FPR at special values are calculated as the quantitative measures for the algorithms. Different colors represent the samples from different datasets, as shown in legends. Different shapes represent different liveness information: $point$=living, $cross$=spoofing.

References

[1] Shervin Rahimzadeh Arashloo, Josef Kittler, and William Christmas. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE access*, 5:13868–13882, 2017. 1, 2

[2] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *FG*, pages 612–618, 2017. 2, 3, 4

[3] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, pages 1–7, 2012. 1, 2, 3, 4

[4] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE TIFS*, 15:42–55, 2019. 2, 3, 4

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 2, 3, 4

[6] Shan Jia, Xin Li, Chuanbo Hu, Guodong Guo, and Zhengquan Xu. 3d face anti-spoofing with factorized bilinear coding. *IEEE TCSVT*, 31(10):4031–4045, 2020. 2, 3, 4

[7] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *CVPR*, pages 8484–8493, 2020. 3, 4

[8] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE TIFS*, 13(7):1794–1809, 2018. 2, 3, 4

[9] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *WACV*, pages 1179–1187, 2021. 2, 3, 4

[10] Siqi Liu, Pong C Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *ECCV*, pages 85–100, 2016. 2, 3, 4

[11] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398, 2018. 1, 2, 3, 4

[12] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *CVPR*, pages 4680–4689, 2019. 2

[13] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *ICB*, pages 158–165, 2018. 3

[14] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031, 2019. 5

[15] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *AAAI*, pages 11974–11981, 2020. 5

[16] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357, 2021. 3, 4

[17] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4, 8

[18] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE TIFS*, 10(4):746–761, 2015. 1, 2, 3, 4

[19] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In *ECCV*, pages 557–575, 2020. 2

[20] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. *IEEE TPAMI*, 43(9):3005–3023, 2020. 2

[21] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *CVPR*, pages 5295–5305, 2020. 1, 2, 3, 4

[22] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *CVPR*, pages 919–928, 2019. 1, 3, 4

[23] Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *ECCV*, pages 70–85, 2020. 2, 3, 4

[24] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *ICB*, pages 26–31, 2012. 1, 2, 3, 4