Journal of Ambient Intelligence and Humanized Computing (2023) 14:16499–16512
https://doi.org/10.1007/s12652-022-03884-z

ORIGINAL RESEARCH

Face anti-spoofing via conditional adversarial domain generalization

Tijian Cai[1] · Fuchun Chen[1] · Wenxin Liu[1] · Xin Xie[1] · Zunxiong Liu[1]

Received: 25 May 2021 / Accepted: 5 April 2022 / Published online: 17 May 2022

Abstract
Many currently existing face anti-spoofing methods do not generalize well to new scenarios due to the changes of background, light, and other factors. To tackle this problem, a face anti-spoofing model based on conditional adversarial domain generalization is proposed in this paper. The model tries to alleviate the discrepancy between source and target domains through the adversarial training of a generator and a domain discriminator. The domain discriminator uses the joint variables generated by multilinear mapping of the features and the classifier predictions as input data. The multiplicative interaction of the input data can promote the domain adversarial model to align multiple domains at the feature and class level, and form a feature space shared by the multiple domains. Besides, the domain discriminator uses the entropy criterion to adjust the priority of samples to reduce the adverse effects of difficult-to-transfer samples with the inaccurate prediction on domain generalization. The generator of the adversarial network consists of attention-Unet and ResNet-18 architectures, where the Unet embedded with the attention mechanism can extract more richer multi-scale domain shared features. The following supervised auxiliary classifier further amplifies the distinguishing features between classes. During the training phase, the model introduces an asymmetric triplet loss in order to get a clearer classification boundary, and introduces a face depth loss to enhance scenario-invariant. Comparative experiments on four public datasets and a custom dataset verify the feasibility of our model. The code is available at https://github.com/17863205785/CADG-master.

Keywords Face anti-spoofing · Conditional adversarial domain generalization · Multi-linear map · Entropy criterion · Face depth

1 Introduction

Face recognition has been widely used as a concerning problem in the field of biometrics (e.g., smartphone unlock, access control, and pay-with-face). However, face presentation attacks (*e.g.*, print attack, video attack, and 3D mask attack) have posed great threats to the security of the face recognition system (Liu et al. 2016, 2018a). To tackle this problem, many face anti-spoofing methods have been proposed and can be roughly divided into machine learning-based methods and deep learning-based methods.

Methods based on traditional machine learning pay more attention to the design of texture features and the use of inherent attributes in images and videos. Most of these methods adopt multi-feature fusion as well as other biological features as auxiliary information to improve the performance, stability, and robustness of the algorithms.

With the success of deep learning, researchers begin to filter high-level semantic features by building a multi-layer convolutional neural network (Krizhevsky et al. 2017). The features learned by deep learning are more discriminative than those extracted by traditional machine learning. But neither machine learning-based methods nor deep learning-based methods can generalize well to new scenarios (Akhtar et al. 2015; Boulkenafet et al. 2017a). To illustrate this problem, a experiment of face anti-spoofing detection using ResNet-18 is done. The features before the output layer

✉ Tijian Cai
lao_cai68@126.com

Fuchun Chen
1577392893@qq.com

Wenxin Liu
1747613113@qq.com

Xin Xie
xienew@qq.com

Zunxiong Liu
153010729@qq.com

[1] Present Address: East China Jiaotong University of China, Nanchang, China

are saved to illustrate the features distribution. The t-SNE (Van der Maaten and Hinton 2008) technology is used to reduce each feature to 2 dimensions, and then the features are displayed on a plane, where each point represents one sample. Fig. 1 shows two cases where the training set and the test set are same or different. Fig. 1a represents the features distribution of intra-dataset testing, both the training set and testing set are CASIA. Figure 1b shows the inter-datasets testing, the training set is CASIA, but the testing set is Replayattack. Figure 1 shows that the performance of the face anti-spoofing can not preserve well in the new application scenarios although the performance of the intra-dataset testing is good. The main reason is that the image texture features, color distortion, and the diversity of attack types cause great differences in the distribution of the features in different domains. To solve this problem, some researchers use domain adaptive methods to align the feature distribution of the source and target domains, thereby improving the generalization performance of the model in the unknown target domain (Damodaran et al. 2018; Hu et al. 2018; Mancini et al. 2018; Pinheiro 2018; Volpi et al. 2018). But most of the target domains are unlabeled and difficult to collect, or there is no information available.

Therefore, researchers began to solve the problem of cross-scene face anti-spoofing from the idea of domain generalization, which trains the model through multiple source domains to make it generalized well to the unknown target domain in the test phase (Ghifary et al. 2015). Enlightening from the Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), the method of adversarial domain generalization is widely used to solve the generalization problem of face anti-spoofing. The domain discriminator is trained to distinguish multiple source domains, while the generator confuses the domain discriminator by learning domain shared features. However, most of the existing adversarial domain generalization methods only align the feature distribution of multiple source domains and ignore the alignment of class level, or

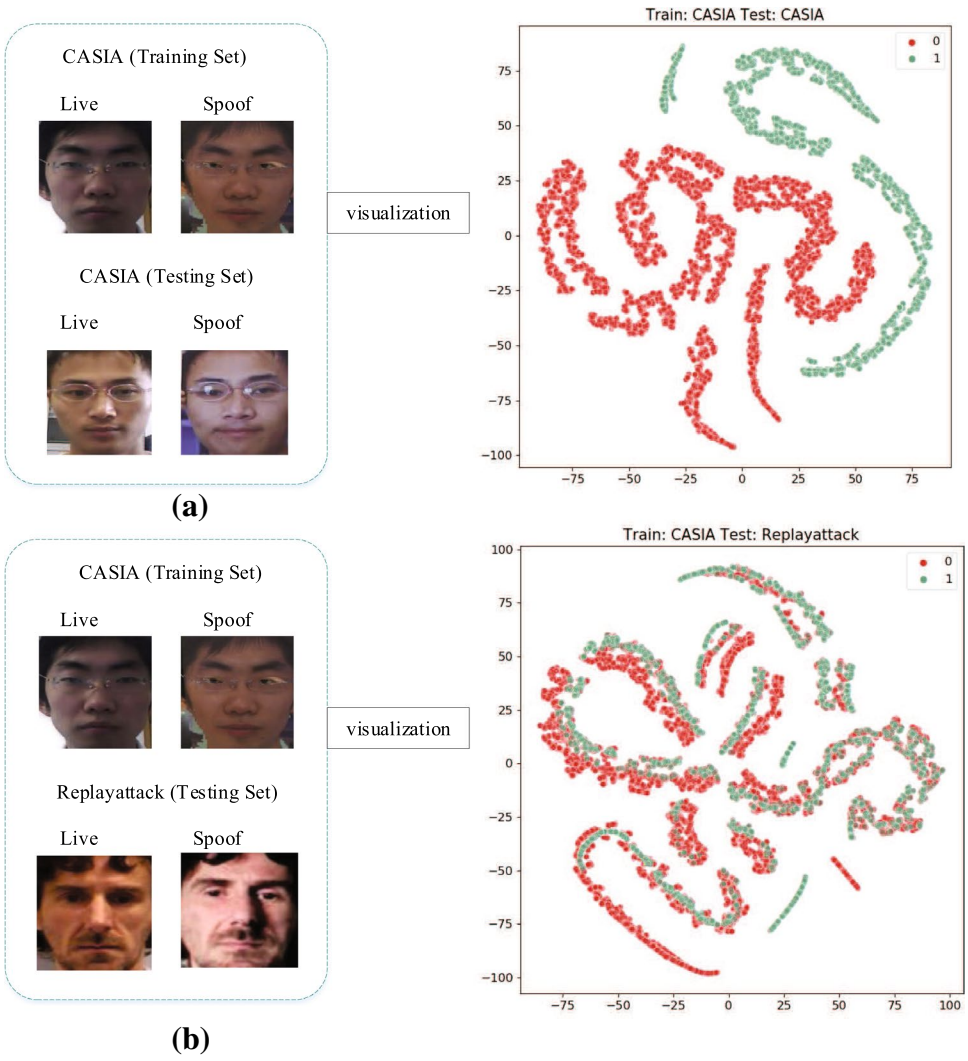


Fig. 1 The T-SNE visualization of the features when training on ResNet-18

align the feature and class distribution separately (Shao et al. 2020; Jia et al. 2020). When the data reflects multimodal structures, it is difficult to align these multimodal structures by above methods, because the multimodal structure can only be fully captured by the cross-covariance dependence between features and classes. In addition, when optimizing the model, the domain discriminator places different samples in the same position, however difficult-to-transfer samples with uncertain predictions may have adverse effects on domain generalization. Therefore, it is very necessary to reduce the impact of these difficult-to-transfer samples.

Inspired by Conditional Generative Adversarial Networks (CGANs) (Mirza and Osindero 2014), we propose a face anti-spoofing model of conditional adversarial domain generalization, which uses the discriminable information transferred in classifier predictions to assist adversarial generalization. We improve the domain discriminator module on the condition of domain-specific feature representation and classifier prediction. Through this conditional mechanism, domain invariance can be achieved at the feature and class level simultaneously. To further improve the generalization ability, we use the entropy criterion to measure the uncertainty of classifier predictions, and adjust the sample weights of domain discriminator to alleviate the adverse effects of hard-to-transfer samples.

In addition, due to the clever means of spoof attack, it is difficult for the model to find a clear classification boundary; Also due to the diversity of attack types, a compact feature space for fake faces is difficult to be obtained. Therefore, it is necessary to use metric learning technology to obtain a clearer classification boundary. The contribution of this article mainly includes the following points:

- we propose a face anti-spoofing model based on conditional adversarial domain generalization. Through conditional constraints, the model aligns multiple source domains at the feature and class level simultaneously, and uses the entropy adjustment to alleviate the adverse effects of difficult-to-transfer samples.
- According to the fact that the fake faces in photos or videos have no 3D structure, the model uses the face depth as scenario-invariant auxiliary information to assist the classifier to improve the robustness of anti-spoofing detection.
- To enhance the discrimination of the deep embedding features, an asymmetric triplet loss is used to constrain model training to get a clearer classification boundary, which makes the distribution of positive samples from different domains more aggregated but the negative more dispersed.
- The effectiveness of the proposed model is verified by comparative experiments with existing state-of-the-art models on four public datasets and a custom dataset.

2 Related works

2.1 Face anti-spoofing methods

The face anti-spoofing methods are mainly introduced from two aspects of machine learning and deep learning. Machine learning methods are more focused on the design of texture features and the use of inherent attributes in images and videos. (Smiatacz 2012) calculated the optical flow values generated by face rotation, trained and classified these optical flow values through SVM (Suykens and Vandewalle 1999). Zhang et al. (2012) proposed the method of color texture analysis to detect whether the image was a real face. On the other hand, the deep learning-based methods mainly extract more discriminative deep features by designing specific network structures and adding auxiliary supervision. Yang et al. (2014) first used a convolutional neural network for face anti-spoofing. Xu et al. (2015) adopted Long Shot-Term Memory and CNN to obtain spatial-temporal features for face anti-spoofing. Shao et al. (2017) proposed a 3D mask face anti-spoofing method to learn robust dynamic texture information from fine-grained deep convolution features. Yu et al. (2020) proposed a central differential convolution network, which could extract the features of the pseudo image well and was not easily affected by the image illumination. Liu et al. (2018b) proposed using depth maps and rPPG signals as the supervision information for CNN learning to improve the generalization ability of the model. Wang et al. (2018) proposed time-series depth information that combineed time-series motion and single-frame face depth, and then used it for face live detection. Wang et al. (2020) proposed a new deep supervision architecture, which used Residual Spatial Gradient Block (RSGB) to capture discriminative details and efficiently encoded spatiotemporal information from a sequence of monocular frames through the Spatio-temporal Propagation Module (STPM). Pérez-Cabo et al. (2019) looked at the problem of face anti-spoofing from the perspective of anomaly detection. They designed a new loss called Triplet Focal Loss, which combined triplet loss and focal loss, and used metric-learning to make the features compact within the class, scatter between classes. Feng et al. (2020) adopted a multi-scale triplet metric learning module, and designed a novel regression loss which only performed the supervision on positive samples to learn more discriminative spoofing clue graphs. However, the performances of these methods are prone to be degraded in the cross-datasets test. This is because the above methods are more inclined to extract the clues in the training datasets

that are biased towards specific attack materials or recording environments. Therefore, this paper proposes to capture more generalized differentiation cues from the perspective of adversarial domain generalization.

2.2 Domain generalization methods

Several domain generalization methods have been proposed. Li et al. (2017) designed a low-rank parameterized CNN model for end-to-end domain generalization learning. Shao et al. (2019) combined the learning of a generalized feature space shared by multiple discriminative source domains with dual-force triplet mining constraints to improve the discriminability of feature space. Shao et al. (2020) applied the existing meta-learning algorithm directly to the face anti-spoofing task, and proposed a novel regularized fine-grained meta-learning framework. However, the above domain generalization methods have only feature alignment, and not class level alignment. The most related work in Jia et al. (2020) proposed a single-side adversarial domain generalization face anti-spoofing model, and made the domain discriminator only inseparable from the real face. But it did not take into account the adverse impact of difficult-to-transfer samples on domain generalization. Instead, in this work, we further add the uncertainty of classifier prediction as a constraint condition of the domain discriminator, giving priorities to easy-to-transfer samples.

3 Research methodologies

The architecture of the methodology of this research is shown in Fig. 2. The whole training network includes three sub networks: feature generator, domain discriminator and auxiliary classifier. The feature generator of the adversarial network is composed of Unet and ResNet-18 architectures. The Unet network embedded with attention mechanism can extract more richer multi-scale domain shared features. The output of the generator is weight-normalized and sent to a supervised auxiliary classifier, which can further amplify the distinguishing features. Motivated by CGANs, we introduce some constraint conditions into the discriminator to associate model training.

The testing network includes the feature generator and the auxiliary classifier. Finally, the results of the classifier are used for face anti-spoofing detection.

3.1 Conditional adversarial domain generalization

Due to the changes of background, light, and other factors, the feature vectors describing the related multiple domains are likely related but potentially different; and as such, their covariates have shifted (Zhou et al. 2021). Domain generalization technique can alleviate the discrepancy between source and target domains. Domain generalization assumes that there exists a generalized feature space underlying the multiple source domains and the target domain, on which the learned model from the seen source domains can generalize well to the unseen target

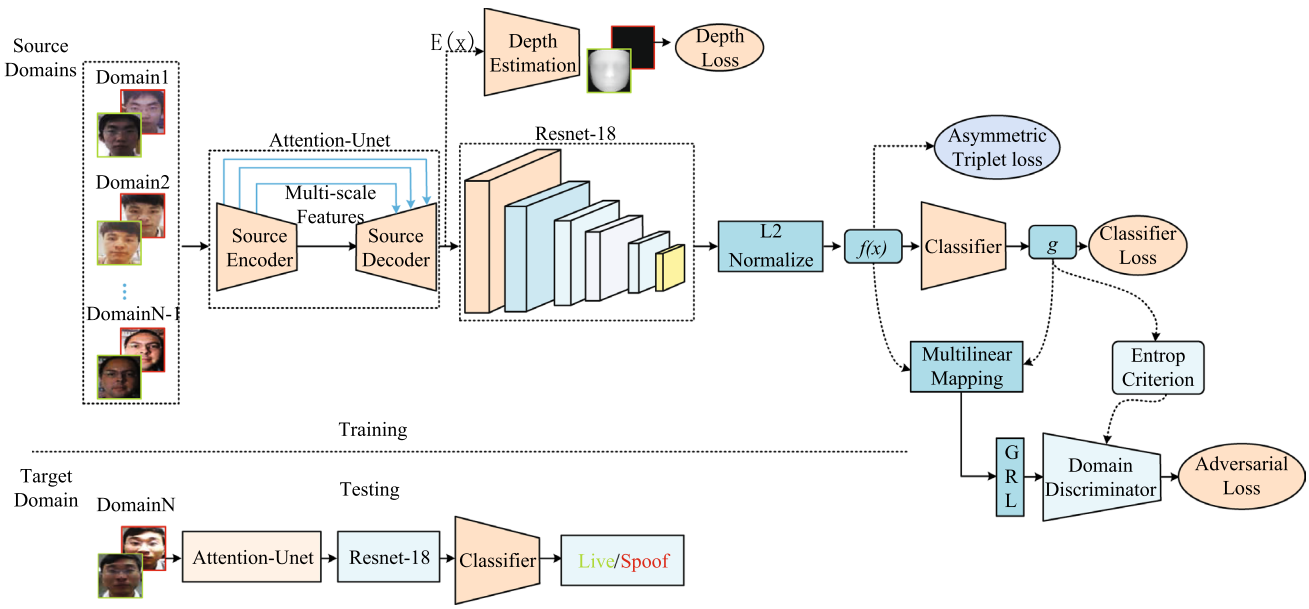


Fig. 2 The overall structure of the model

domain. Suppose there are N source domains, denoted as $D = \{X_1, X_2, ..., X_N\}$, the domain labels are defined as $Y = \{Y_1, Y_2, ..., Y_N\}$, x represents the input sample from X, y is the domain label of x. Each domain contains two categories of face images, the real face and the fake face. Therefore, for the discriminator, it is necessary to identify $2N$ categories, that is, distinguish not only which source domain the sample comes from, but also which category the sample belongs to. The training of domain adversarial network is a minimax optimization problem as follows:

$$f = G(x), h = (f, g) \tag{1}$$

$$\min_D \max_G L_{Ada}(G, D) = -E_{x,y\sim X,Y} \sum_{n=1}^{N} \mathbb{1}_{[n=y]} \log D(h) \tag{2}$$

where f is the feature extracted by the feature extractor G, and g is the classifier prediction, $h = (f, g)$ represents the joint variable of f and g, and is the input data of discriminator D. $\mathbb{1}$ is the indicator function. When n=y, that is, when the discriminator D judges which domain the sample belongs to correctly, the indicator function is 1, otherwise it is 0. L_{Ada} represents the loss of adversarial training. The generator G is trained for maximizing the adversarial loss, while domain discriminator D is optimized in the opposite direction. Through the adversarial training of the generator and the domain discriminator, the domain adversarial network intends to alleviate the discrepancy between multiple domains in order to improve the generalization performance of the model.

Moreover, we add a gradient reversal layer (GRL) before the domain discriminator. It means that the gradient of the generator will be multiplied by $-\lambda$ in the backpropagation process. We set $\lambda = \frac{2}{1+\exp(-10k)} - 1$, where $k = \frac{\text{current\_iters}}{\text{total\_iters}}$, current_iters is the number of current iterations, total_iters is the total number of iterations. In this way, we can optimize the feature generator as well as domain discriminator simultaneously, which can reduce the complexity of adversarial training.

3.1.1 Multi-linear mapping

Motivated by CGANs, we introduce some constraint conditions to discriminator to associate model training. Firstly, in order to make full use of the multimodal information in the classifier prediction g and better represent the multiplicative interaction between the feature f and the classifier prediction g, we use multi-linear mapping (Zhao et al. 2021) to combine f and g. The multi-linear mapping is defined as the outer product of multiple random vectors, $T_{\otimes}(h) = f \otimes g$, where $\otimes$ represents the outer product. Let d_f and d_g denote the dimensions of vectors f and g, respectively. The multi-linear mapping has a dimension of $d_f \times d_g$, which is often too high, and it is likely to cause dimension explosion. In this work, we solve the dimension explosion by random sampling strategy proposed in Laparra et al. (2015), Kar and Karnick (2012). The idea is to approximate the outer -product $T_{\otimes}(h) = f \otimes g$ using the dot-product:

$$T_{\odot}(f, g) = \frac{1}{\sqrt{d}}(R_f f) \odot (R_g g) \tag{3}$$

where $\odot$ is the element-wise product. $R_f \in R^{d\times d_f}$ and $R_g \in R^{d\times d_g}$ are two random matrices sampled only once and fixed in the training phase. d is a hyperparameter that represents the dimension to be sampled, usually $d \ll d_f \times d_g$, $R_f f$ and $R_g g$ are randomly sampled vectors of f and g, which have the same dimension d. In this way, the data dimension after fusion will not be very large.

Through the adversarial training of the joint variables, the model can align multiple source domains at the feature and class levels simultaneously, which can fully align the data distribution of multiple source domains.

3.1.2 Entropy adjustment

Generally, domain discriminator puts different samples on the same importance. However, because hard-to-transfer samples with uncertain predictions may adversely affect domain generalization, it is necessary to reduce their weights to weaken their impacts. In our model, we adopt the entropy criterion $H(g) = -\sum_{c=1}^{C} g_c \log g_c$ to quantify the uncertainty of the classifier prediction, where C is the number of categories, and g_c is the probability of the sample belong to class c. Then each sample can obtain an entropy-aware weight $\omega(H(g)) = 1 + e^{-H(g)}$,which is used to re-weight the training sample. When the uncertainty of classifier prediction is larger, the ω is smaller, thereby the bad impact of hard-to-transfer samples on domain generalization will be weakened. After entropy adjustment, the objective function of the optimization of the conditional adversarial is:

$$\min_D \max_G L_{Ada}(G, D) = -E_{x,y\sim X,Y}\omega(H(g)) \sum_{n=1}^{N} \mathbb{1}_{[n=y]} \log D(h) \tag{4}$$

3.2 Asymmetric triplet loss

Due to the clever means of spoof attack, it is difficult for the model to find a clear classification boundary. Moreover, due to the complex attack types, such as photo attack, video replay attack, 3D mask attack, etc., the feature distribution discrepancies of fake faces are large, and it is difficult to find

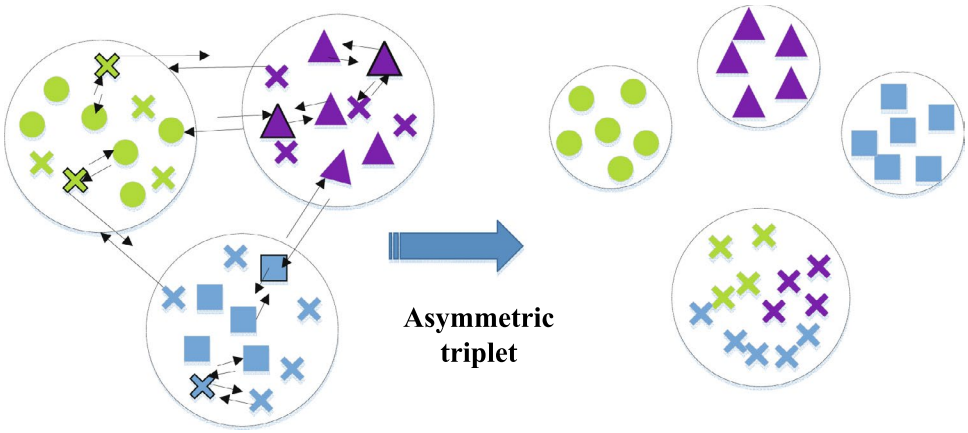


Fig. 3 Mining hard negative samples in asymmetric triple loss

a compact feature space for fake faces. For this reason, we adopt an asymmetric triplet loss (Kertész 2021) constraint to make the distribution of fake faces in different domains dispersed, while the real faces in different domains compact. The optimization key of asymmetric triplet loss is mining approach of hard samples. Unlike triplet loss, asymmetric triplet loss mines hard samples from multiple domains, that is to say, one mini-batch contains the samples of multiple domains. As shown in Fig. 3, circles, squares, and triangles represent the attack samples from different domains, while crosses represent real faces in different domains; the samples marked with a black border represent the anchor samples. In order to make the real faces in different domains compact, the anchor samples and the positive samples can be real faces from different domains; In order to make the real faces and fake faces in same domains dispersed, the anchor samples and the negative samples can be real faces and fake faces from same domains.

Assuming there are N source domains, the real and the fake faces are recombined into $N+1$ categories. The fake faces that come from N source domains are considered as N distinct categories, but all the real faces are treated as one category. By minimizing the asymmetric triplet loss constraint, the real faces and the fake faces are separated, and $N+1$-categories asymmetric triplet loss constraint is as follows:

$$\min L_{Triplet}(G) = \sum_{x^a x^p x^n}^{G} (\|f(x^a) - f(x^p)\|_2^2 - \|f(x^a) - f(x^n)\|_2^2 + \alpha) \tag{5}$$

Where $f(x)$ represents the feature of sample x, x^a, x^p, x^n are anchor sample, positive sample and negative sample respectively. x^a and x^p are the same category, but x^a and x^n are different. α is the threshold, which controls the distance between positive samples and negative samples.

3.3 Face depth auxiliary supervision

In order to further improve the generalization ability of the model, the face depth map is used as a scenario-invariant auxiliary supervision. Since the real face have certain depth among the nose, mouth, and forehead, in other words, there are 3D depth structure information in real faces. However, the fake faces in photos or videos have no 3D structure. For these reasons, we train a depth estimator in a supervised manner to improve the model. The training target values of depth maps of the fake faces are set to 0 since the fake faces are flat; while the 3D facial structures of the real faces are reconstructed from single 2D images (Feng et al. 2018) by a face alignment network named PRNet. These estimated 3D depth values will be used as the training target values of the real faces. As shown in Fig. 4, there are some real faces and fake faces from four public datasets. The second line

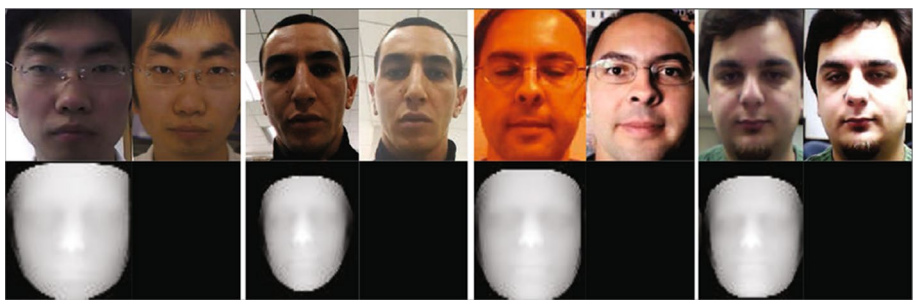


Fig. 4 Face depth map estimated by PRNet in four public datasets

is the corresponding depth map. By minimizing the depth loss constraint, real face and fake face can be further distinguished. Face depth auxiliary supervision is defined as:

$$L_{Dep}(X;Dep) = \|Dep(E(x)) - I(x)\|_2^2 \quad (6)$$

Where $E(x)$ is the output of attention-Unet, $Dep(E(x))$ represents the depth map estimator, and $I(x)$ is the assumed truth map estimated by PRNet.

3.4 Loss function

In our model, a supervised auxiliary classifier is used to amplify the distinguishing features. The features extracted by the generator are easily affected by the environmental factors (Ranjan et al. 2017), such as illumination, camera resolution, etc., which make great differences in feature norms across scenes. Therefore, l_2 normalization is performed on the output of the feature encoder and on the weights of the classification layer to restrict them to share the same Euclidean norm.

The classifier uses cross entropy as the objective function. Therefore, based on the above research, the overall optimization objective of the model is as follows:

$$L_{DG} = \lambda_1 L_{Cls} + \lambda_2 L_{Ada} + \lambda_3 L_{Dep} + \lambda_4 L_{Triplet} \quad (7)$$

where L_{Cls} is the cross-entropy loss of the classifier. L_{Ada} is the conditional adversarial loss. L_{Dep} is the regression loss of depth map. $L_{Triplet}$ is the asymmetric triplet loss. $\lambda_1 \sim \lambda_4$ are hyperparameters which can balance the influence of the four loss functions on the whole model. The hyperparameters are initialized according to (Shao et al. 2019) and (Jia et al. 2020), etc., and fine tuned according to the contributions of the four loss functions in ablation experiments with the development sets. Through the adversarial training under the constraints of the four loss functions, we can generate a more generalized domain shared feature space, so that the model can better generalize to unseen domains.

4 Experiments

4.1 Experimental setting

Datasets. Four public FAS datasets are used to evaluate the performance of the proposed model: CASIA-MFSD (Zhang et al. 2012) (abbreviated as C), Oulu-NPU (Boulkenafet et al. 2017b) (abbreviated as O), Replay-Attack (Chingovska et al. 2012) (abbreviated as I), MSU-MFSD (Wen et al. 2015) (abbreviated as M). The CASIA includes three different types of attacks: bending photo, cropping the photo and video attacks; the Oulu contains two types of attacks: photo and video replay attacks; Replayattack includes three types of attacks: print, mobile, and highdef attack; MSU mainly contains two different spoofing attacks: printed photo and video replay attack. During training, we randomly select three datasets from the four datasets as the source domain, and the remaining one is regarded as the unseen target domain for the test phase, and there are four scenarios for

Table 1 The structure details of all modules of the proposed network

Feature generator			Discriminator			Feature embedder			Depth estimator		
Layer	Chan./stri	Out.size	Layer	Chan./stri	Out.size	Layer	Chan./stri	Out.size	Layer	Chan./stri	Out.size
Input image			Input fc3-1			Input conv1-4			Input conv1-4		
Conv1/layer0	3/1	128	Fc2-1		512	Conv3-1	3/2	64	Conv4-1	3/2	64
Max-pool1-1	-/2	64	Fc2-2		3	Max-pool3-1	-/2	32	Conv4-2	128/2	32
Encoder-2	64/-	32				Layer2	64/-	16	Conv4-3	64/1	32
Encoder-3	256/-	16				Layer3	128/-	8			
Encoder-4	512/-	8				Layer4	256/-	4			
Encoder-5	1024/-	4				AdaptiveAvgpool		1			
Center	2048/-	4				Fc3-1		512			
Decoder-5	4096/-	8									
Decoder-4	1024+32/-	16									
Decoder-3	512+32/-	32									
Decoder-2	256+32/-	64									
Decoder-1	32/-	128									
Conv1-2	160/1	128									
Conv1-3	64/1	128									
Conv1-4	32/1	128									

inter-datasets testing: I&C&Mto O, O&C&I to M, O&C&M to I, and O&M&I to C. Affected by the types of attack, the scenes of the four datasets are very different, which brings great challenges to cross-scene testing.

The structure of network. The specific structure of the network implemented by PyTorch is shown in Table 1. Attention-Unet is used as a feature generator to extract multi-scale features. The SE-ResNet-50 (Hu et al. 2020) is selected as the pre-training model and the layer0∼layer4 of SE-ResNet-50 is used as the encoder. In the decoder part, the spatial attention-mechanism is added into the upsampling process to extract domain shared space features and face region features. The results of each upsampling are converted to the same size through the bilinear interpolation, and concatenated as the input of the feature embedder. The layer2∼layer4 of ResNet-18 (He et al. 2016) are used as the network structure of feature embedder and then an adaptive average pooling layer and a fully connected layer (FC) with 512 hidden nodes are designed later. The classifier contains two FC layers with 512 and 2 nodes. The domain discriminator takes the joint variables generated by multilinear mapping as input data. It contains two FC layers with 512 and 3 nodes, respectively. The depth estimator contains three convolution layers, a batch normaliza-tion layer and a rectified linear unit activation function(RELU). Its convolutional kernel size is set to 3×3.

Evaluation Metrics. The most common metrics are used in both intra and cross-testing experiments, including Area Under the Curve (AUC), Half Total Error Rate (HTER), and Accuracy (Acc). HTER is found out by calculating the average of FRR (ratio of incorrectly rejected bonafide score) and FAR (ratio of incorrectly accepted attacks). AUC represents the degree of separability between bonafide and spoofings. Acc is used to measure the correct proportion of sample classification.

Implementation Details. The face detection and alignment algorithm MTCNN is used (Zhang et al. 2016) for data pre-processing. All RGB face images are cropped into $256 \times 256 \times 3$, resized to $128 \times 128 \times 3$,and then augmented. Each video randomly selects a frame as the input of feature generator. The SGD is selected as the optimizer, with momentum set to 0.9, weight decay 5e-4, learning rate initialized to 0.001 and dropped to 0.1 times of the previous per 100 epochs.

During the training phase, we use an end-to-end approach to train the model. The batch-size of each domain is 20, so a total of 60 for the 3 domains. The hyper-parameters $\lambda_1 \sim \lambda_4$ are set to 1.0, 0.5, 0.5, 1.0. During the testing phase, we randomly select two frames from each video in the target domain as input data. The anti-spoofing detection is determined according to the results of the classifier.

4.2 Comparison with Baseline model

To verify the performance of our model, we first compare our model with some common face anti-spoofing models on four testing tasks (I&C&Mto O, O&C&I to M, O&C&M to I, and O&M&I to C). The comparison models include MS-LBP (Määttä et al. 2011); Binary CNN (Yang et al. 2014); IDA (Wen et al. 2015); Color Texture (Boulkenafet et al. 2016); LBP-TOP (de Freitas Pereira et al. 2014); Auxiliary (Liu *et al.* 2018). The experimental results are shown in Table 2. The HTER of our model is the lowest, and the AUC of our model is the highest, which prove our model outperforms the above state-of-the-art models. Most state-of-the-art models can perform well in intra-dataset testing, but the generalization ability degrades when testing in new scenarios. The main reason is that these models do not fully consider the internal correlation of the data distribution among multiple domains, and the extracted features are mostly domain-specific features.

Secondly, we compare our model with some state-of-the-art domain generalization models in the face anti-spoofing task. The experiment results are shown in Table 3. It can be seen from the results that our model outperforms MMD-AAE, MADDG (Shao et al. 2019), SSDG-M (Jia et al. 2020), and SSDG-R. The model of MADDG extracts domain sharing features through the idea of multi-adversarial and dual-force triplet mining constraints, but it only aligns features and ignores the alignment of class level. The SSDG model takes

Table 2 Comparison of performance with common face anti-spoofing models on four testing sets

Model	I&C&M to O		O&C&I to M		O&M&I to C		O&C&M to I	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MS-LBP	50.3	49.3	29.8	78.5	54.3	45.0	50.3	51.6
Binary CNN	29.6	77.5	29.3	82.9	34.9	72.0	34.5	65.9
IDA	54.2	44.6	66.7	27.9	55.2	39.1	28.4	78.3
Color Texture	63.6	32.7	28.1	78.5	30.6	76.9	40.4	62.8
LBP-TOP	53.2	44.1	37.0	70.8	42.6	61.1	49.5	49.5
Auxiliary(Depth)	30.2	77.6	22.7	85.9	33.5	73.2	29.1	71.7
Auxiliary	–	–	-	-	28.4	–	27.6	–
Ours	13.5	92.1	1.3	99.9	10.7	95.7	11.1	95.8

Table 3 Comparison of performance with domain generalization face anti-spoofing models on four testing sets

Model	I&C&M to O		O&C&I to M		O&M&I to C		O&C&M to I	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MMD-AAE	40.9	63.1	27.1	83.2	44.6	58.3	31.6	75.2
MADDG	27.9	80.0	17.7	88.1	24.5	84.5	21.2	85.0
SSDG-M	25.5	80.8	2.4	99.8	24.8	80.8	15.8	90.7
SSDG-R	19.4	90.7	2.4	99.1	21.0	83.3	14.3	90.1
Ours	13.5	92.1	1.3	99.9	10.7	95.7	11.1	95.8

into account the diversity of the fake faces, so it proposes a single-side adversarial framework and finds a domain-invariant feature space for the real faces. However, it does not require class level alignment, and ignores the adverse effect of hard-to-transfer samples on domain generalization. Our conditional adversarial domain generalization model adds the multi-modal category information into adversarial training, which makes the multiple source domains align at the feature and class level simultaneously. The entropy adjustment effectively reduces the adverse effects of hard-to-transfer samples.

From the comparison experiment, we can see that our model can obtain the best accuracy metrics. However, the average test frequency of our network is about 26.4 fps, which is lower than 36.4 fps of MADDG algorithm. This running speed is not ideal, but it is within the acceptable range. The test network includes the attention-UNet, the Resnet-18 and the classifier, which is much simpler than the training network. However, it still needs to be optimized by pruning technology, etc., which is beyond the scope of this paper.

4.3 Custom dataset experiment

To verify the performance of our model in ground truth environment, we use ordinary mobile cameras to collect videos and generate the samples of photo attack and video replay attack to establish a custom dataset. Some real faces and fake faces are shown in the third and fourth lines of the Fig. 7. The custom dataset consists of short video with resolution (640 by 480 pixels) recordings of 24 different identities, and adopts various video acquisition schemes: (1) the background of the scene is uniform or non-uniform; (2) the operator holds the attack device using their own hands or sets the attack devices on a fixed support; (3) the operator displays the videos using an iPhone screen or using an iPad screen.

Imitating the above comparison experiments, we use the custom dataset as the test set and design four scenarios for inter-datasets testing: I&M&C to U, O&C&I to U, O&I&M to U, O&C&M to U, where U represents the custom dataset. The results are shown in Table 4. From Table 4, we can see that the experimental results on the custom dataset are similar to those on public datasets. This shows that the experimental results of our model are reliable and stable.

4.4 Discussion

4.4.1 Ablation Study

We design the ablation experiments to verify the feasibility of the modules in our framework of conditional adversarial domain generalization. The corresponding modules are defined as 'w/o attention', 'w/o triplet', 'w/o adversarial', 'w/o multilinear & entropy', 'w/o norm'. The 'w/o attention' means removing the attention mechanism from the Unet feature extractor. The 'w/o triplet' means removing the asymmetric triplet constraint. The 'w/o adversarial' means removing conditional adversarial domain discriminator. The 'w/o multilinear & entropy' means that the network contains the domain discriminator but does not add multilinear mapping and entropy criterion. The 'w/o norm' denotes that the proposed network does not include feature and weight normalization module. The ablation results of above modules under inter-datasets testing on CASIA, OULU, MSU, and Repalysttack are shown in Table 5, and the corresponding ROC curves are shown in Fig. 5.

From Table 5, we can see that the performance of the proposed model degrades when any modules are excluded. This indicates that the six modules are beneficial to the whole model. Moreover, we find that conditional adversarial domain discriminator and asymmetric triplet constraint have a greater impact on the experiment results than other modules. For example, in the inter-datasets testing on CASIA and Replayattack, removing the conditional adversarial domain

Table 4 Evaluation of our model on the custom dataset

evaluation metrics	I&M&C to U	O&C&I to U	O&I&M to U	O&C&M to U
Acc(%)	93.24	85.02	88.89	82.60
HTER(%)	7.63	7.27	5.82	14.92
AUC(%)	97.77	96.91	97.05	93.55

Table 5 Evaluation of the influence of each module in ablation experiments on four datasets

model	I&C&M to O		O&C&I to M		O&M&I to C		O&C&M to I	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
w/o attention	17.5	87.1	2.9	97.2	18.7	89.0	17.2	89.9
w/o triplet	20.9	87.3	5.7	97.8	24.7	81.3	22.6	86.6
w/o adversarial	16.5	88.9	2.8	98.2	20.6	85.5	21.4	83.0
w/o norm	22.0	84.1	0.8	99.0	17.4	92.3	17.9	90.6
w/o depth	16.1	92.3	2.7	99.6	16.7	90.8	16.7	90.1
w/o multilinear&entropy	15.5	90.4	2.1	99.9	17.4	89.9	17.1	91.4
Ours	13.5	92.1	1.3	99.9	10.7	95.7	11.1	95.8

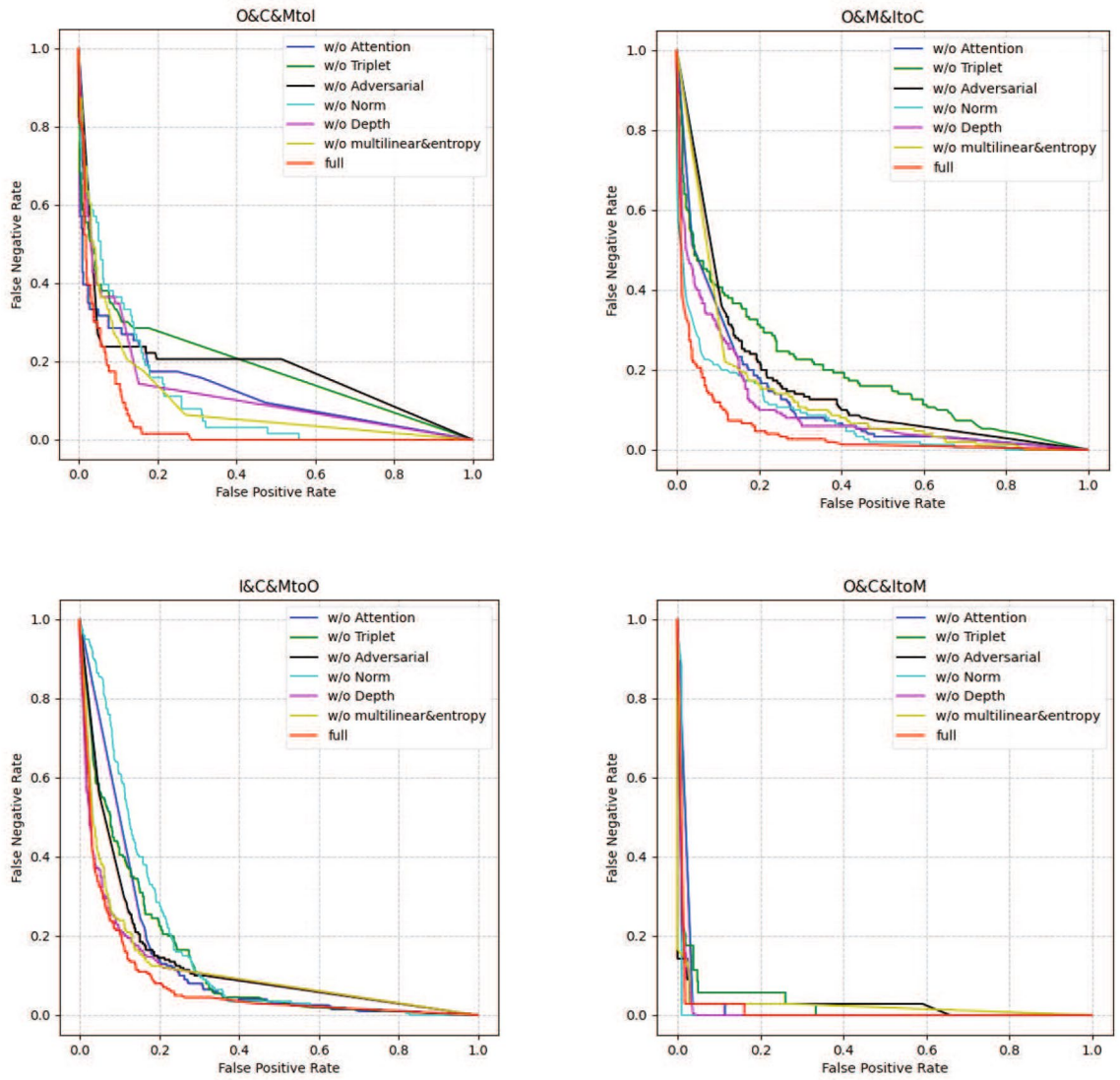


Fig. 5 ROC curves of ablation learning on four public datasets

discriminator will cause HTER to rise by 9.9% and 10.3%, respectively, removing the asymmetric triplet constraint from our model leads to 14.0% and 11.5% higher HTER. The main reason is that our model not only aligns the features of multiple source domains, but also aligns the predictions of the classifier, thus forming a more generalized domain-invariant feature

space. From Fig. 5, we can directly see the impact of each module on the evaluation metric of ROC. The curves of 'w/o triplet' anf 'w/o adversarial' are generally above other curves, which shows that they have a greater impact on the model.

Figure 6 is t-SNE visualization of the feature space when the target domain is Replayattack. Figure 6a removes the multi-linear map & entropy adjustment; Fig. 6(b) removes the conditional adversarial domain discriminator module; Fig. 6c removes the asymmetric triplet constraint; Fig. 6(d) contains all the modules. We can see that the classification boundary of Fig. 6d is the clearest among the figures. This means that all modules are helpful to improve the generalization ability of the model in cross-scene face anti-spoofing, so it proves the feasibility of our model.

4.4.2 Limited source domains

In this experiment, we verify the effect of the number of source domains on the generalization performance of the model. We limit the number of source domains to 2 datasets during training. Then its testing performance is compared with those of 3 datasets. Due to significant domain-variation features that exist in MSU and Replayattck, these two datasets are selected as source domains for training, and OULU as well as CASIA are selected as the target domains for testing. The experiment results are shown in Table 6. We can see that our model has better results than other models, such as MS-LBP, IDA, Color Texture, LBP-TOP, MADDG, SSDG-M. Comparing the data in Table 2, we can see that our model can learn more generalization clues when increasing the number of source domains. However, the other models have little improvement when the number of training domains is increased.

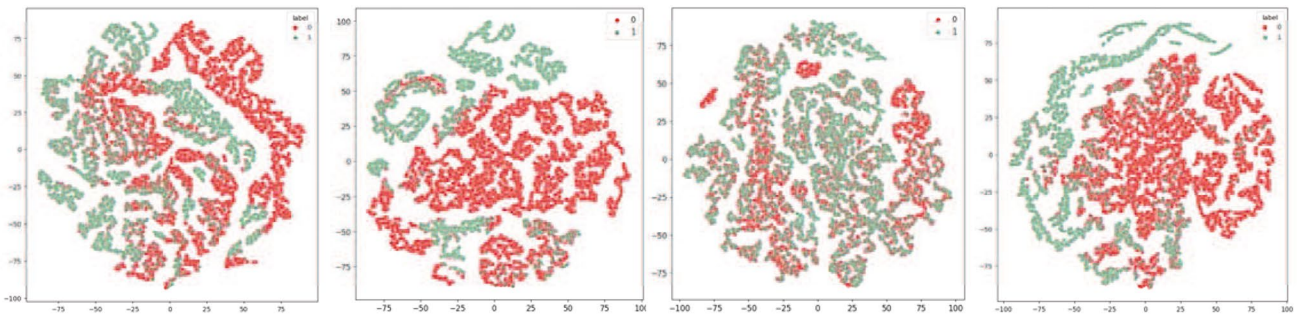
Table 6 The comparison results of domain generalization face anti-spoofing under limited source domains

Model	M&I to C		M&I to O	
	HTER	AUC	HTER	AUC
MS-LBP	51.16	52.09	43.63	58.07
IDA	45.16	58.80	54.52	42.17
Color Texture	55.17	46.89	53.31	45.16
LBP-TOP	45.27	54.88	47.26	50.21
MADDG	41.02	64.33	39.35	65.1
SSDG-M	31.89	71.29	36.01	66.88
Ours	27.33	78.90	30.44	75.58

4.5 Visualizations of Proposed model

Class Activation Mapping (CAM) is usually used to visualize deep learning features. It can locates key parts of the image through feature response, and provides a model for deep learning interpretability. CAM displays the strength information of the image's local response in the form of a heat map. The local region with stronger response has better feature recognition abilities. We use the networks trained with O&I&M to generate feature maps and weights, and use the Grad-CAM(Selvaraju et al. 2017) to provide the class activation map (CAM) visualization of our model. The Grad-CAM are shown in Fig. 7, the first line and the second line are real faces and fake faces of CASIA respectively, the third line and the fourth line are real faces and fake faces of the custom dataset. Our conditional adversarial domain generalization model pays more attention to the features of the facial region (such as eyes and nose region) rather than the background, lighting, etc., of different domains.

In addition, in order to show the process of training, we use t-SNE to visualize the changes of the features distribution with iterative training of O&C&I to M task. As shown in Fig. 8, where domain1, domain2 and domain3 represent the three source domains CASIA, OULU, and Replyattack



(a) w/o multilinear&entropy (b) w/o adversarial (c) w/o triplet (d) full

Fig. 6 The t-SNE visualization of feature space for ablation learning on O&C&M to I test task

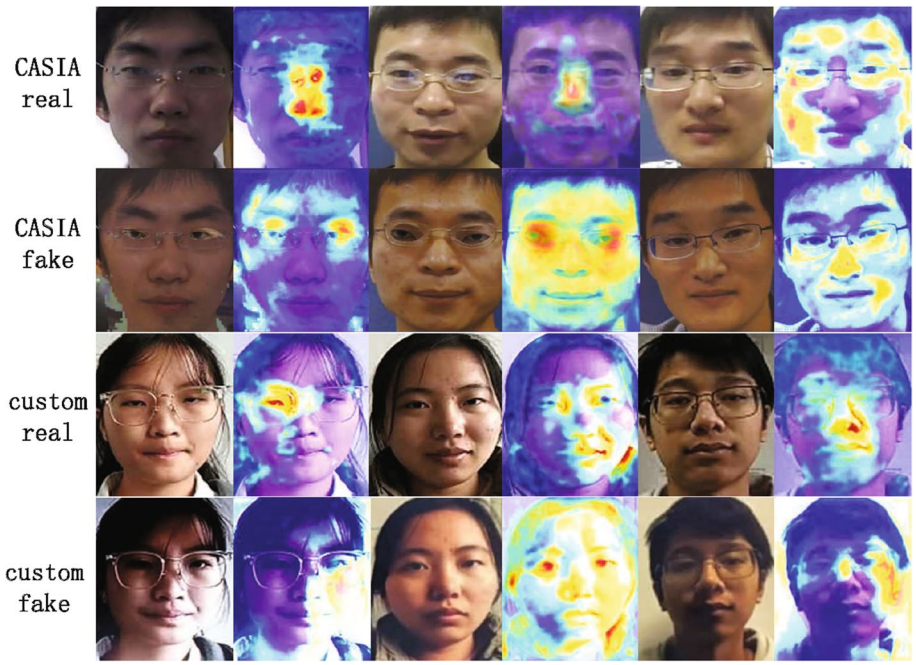


Fig. 7 Grad-CAM visualization of the network trained with O&I&M

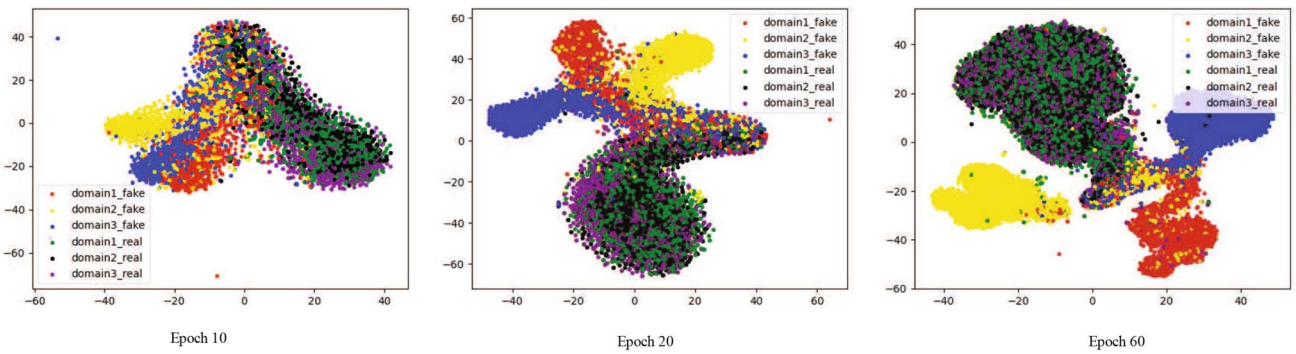


Fig. 8 The t-SNE visualization of the feature changes on C&O&I to M training task

respectively. With the increase of epoch numbers, the positive samples of different domains are more compact, while the negative samples are more dispersed. The classification boundary is gradually clear. At the same time, the discrepancy among multiple domains is gradually disappearing.

5 Conclusion

In this study, we proposed a conditional adversarial domain generalization model to improve the generalization ability for face anti-spoofing. The feasibility of our model is verified by the experiments on four public anti-spoofing datasets and the custom dataset. The experiments confirm that adversarial training with joint variables can alleviate the discrepancy between source and target domains, promoting the model to align multiple source domains at the feature and class level simultaneously. The entropy adjustment can reduce the adverse effects of the samples of inaccurate prediction. The experiments also confirm that the asymmetric triplet loss constraint can promote the fake face in different domains more separated while keeping the real ones aggregated, and face depth loss constraint can further improve the performance of face anti-spoofing detection of photo and video attack types. Although our model outperforms several state-of-the-art models on accuracy metrics, the test speed of the model is a little slower than that of MADDG model. This can be improved by using lightweight modules or pruning technology, which will be our later works. In addition, the mining of asymmetric triplet samples is limited to a single mini-batch. Current state-of-the-art technology of

Cross-Batch Memory can break the limit. Later, we will work on these aspects to further improve our model.

Acknowledgements This work is supported by Key Research and Development Program of Jiangxi Province (Grant no. 20203BBE53029 and Grant no. 20202BBEL53004). We would also like to thank our team of Deep Data Science for the valuable contributions.

Declarations

Conflict of interest The authors declare that there is no conflict of interests in the paper.

References

Akhtar Z, Micheloni C, Foresti GL (2015) Biometric liveness detection: challenges and research opportunities. IEEE Secur Priv 13(5):63–72. https://doi.org/10.1109/MSP.2015.116

Boulkenafet Z, Komulainen J, Hadid A (2016) Face spoofing detection using colour texture analysis. IEEE Trans Inf Forensics Secur 11(8):1818–1830. https://doi.org/10.1109/TIFS.2016.2555286

Boulkenafet Z, Komulainen J, Hadid A (2017a) Face antispoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Process Lett 24(2):141–145. https://doi.org/10.1109/LSP.2016.2630740

Boulkenafet Z, Komulainen J, Li L, Feng X, Hadid A (2017b) Oulunpu: a mobile face presentation attack database with real-world variations. IEEE Int Conf Autom Face Gesture Recogn 5:5. https://doi.org/10.1109/FG.2017.77

Chingovska I, Anjos A, Marcel S (2012) On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG), IEEE, pp 1–7

Damodaran BB, Kellenberger B, Flamary R, Tuia D, Courty N (2018) Deepjdot: deep joint distribution optimal transport for unsupervised domain adaptation. Proc Eur Conf Comput Vis (ECCV). https://doi.org/10.1007/978-3-030-01225-0_28

de Freitas Pereira T, Komulainen J, Anjos A, De Martino JM, Hadid A, Pietikäinen M, Marcel S (2014) Face liveness detection using dynamic texture. EURASIP J Image Video Process 1:1–15. https://doi.org/10.1186/1687-5281-2014-2

Feng Y, Wu F, Shao X, Wang Y, Zhou X (2018) Joint 3d face reconstruction and dense alignment with position map regression network. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision – ECCV 2018. Lecture notes in computer science, vol 11218. Springer, Cham. https://doi.org/10.1007/978-3-030-01264-9_33

Feng H, Hong Z, Yue H, Chen Y, Wang K, Han J, Liu J, Ding E (2020) Learning generalized spoof cues for face anti-spoofing. arXiv:2005.03922

Ghifary M, Kleijn WB, Zhang M, Balduzzi D (2015) Domain generalization for object recognition with multi-task autoencoders. IEEE Int Conf Comput Vis (ICCV). https://doi.org/10.1109/ICCV.2015.293

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. arXiv:1406.2661

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proc IEEE Conf Comput Vis Pattern Recogn. https://doi.org/10.1109/CVPR.2016.90

Hu L, Kan M, Shan S, Chen X (2018) Duplex generative adversarial network for unsupervised domain adaptation. IEEE/CVF Conf Comput Vis Pattern Recogn. https://doi.org/10.1109/CVPR.2018.00162

Hu J, Shen L, Albanie S, Sun G, Wu E (2020) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42(8):2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372

Jia Y, Zhang J, Shan S, Chen X (2020) Single-side domain generalization for face anti-spoofing. IEEE/CVF Conf Comput Vis Pattern Recogn (CVPR). https://doi.org/10.1109/CVPR42600.2020.00851

Kar P, Karnick H (2012) Random feature maps for dot product kernels. In: Lawrence ND, Girolami M (eds) Proceedings of the fifteenth international conference on artificial intelligence and statistics. PMLR, La Palma, pp 583–591

Kertész G (2021) Different triplet sampling techniques for lossless triplet loss on metric similarity learning. IEEE World Symp Appl Mach Intell Inf (SAMI). https://doi.org/10.1109/SAMI50585.2021.9378628

Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 60(6):84–90. https://doi.org/10.1145/3065386

Laparra V, Gonzalez DM, Tuia D, Camps-Valls G (2015) Large-scale random features for kernel regression. IEEE Int Geosci Remote Sens Symp (IGARSS). https://doi.org/10.1109/IGARSS.2015.7325686

Li D, Yang Y, Song YZ, Hospedales TM (2017) Deeper, broader and artier domain generalization. IEEE International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2017.591

Liu S, Yuen PC, Zhang S, Zhao G (2016) 3d mask face anti-spoofing with remote photoplethysmography. European conference on computer vision. Springer, Berlin, pp 85–100. https://doi.org/10.1007/978-3-319-46478-7_6

Liu SQ, Lan X, Yuen PC (2018a) Remote photoplethysmography correspondence feature for 3D mask face presentation attack detection. Proc Eur Conf Comput Vis (ECCV). https://doi.org/10.1007/978-3-030-01270-0_34

Liu Y, Jourabloo A, Liu X (2018b) Learning deep models for face anti-spoofing: binary or auxiliary supervision. IEEE/CVF Conf Comput Vis Pattern Recogn. https://doi.org/10.1109/CVPR.2018.00048

Määttä J, Hadid A, Pietikäinen M (2011) Face spoofing detection from single images using micro-texture analysis. Int Jt Conf Biom (IJCB). https://doi.org/10.1109/IJCB.2011.6117510

Mancini M, Porzi L, Bulo SR, Caputo B, Ricci E (2018) Boosting domain adaptation by discovering latent domains. IEEE/CVF Conf Comput Vis Pattern Recogn. https://doi.org/10.1109/CVPR.2018.00397

Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv:1411.1784

Pérez-Cabo D, Jiménez-Cabello D, Costa-Pazo A, López-Sastre RJ (2019) Deep anomaly detection for generalized face anti-spoofing. IEEE/CVF Conf Comput Vis Pattern Recogn Worksh (CVPRW). https://doi.org/10.1109/CVPRW.2019.00201

Pinheiro PO (2018) Unsupervised domain adaptation with similarity learning. IEEE/CVF Conf Comput Vis Pattern Recogn. https://doi.org/10.1109/CVPR.2018.00835

Ranjan R, Castillo CD, Chellappa R (2017) L2-constrained softmax loss for discriminative face verification. arXiv:1703.09507

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. IEEE Int Conf Comput Vis (ICCV). https://doi.org/10.1109/ICCV.2017.74

Shao R, Lan X, Yuen PC (2017) Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3D mask face anti-spoofing. IEEE Int Jt Conf Biom (IJCB). https://doi.org/10.1109/BTAS.2017.8272765

Shao R, Lan X, Li J, Yuen PC (2019) Multi-adversarial discriminative deep domain generalization for face presentation attack detection. IEEE/CVF Conf Comput Vis Pattern Recogn (CVPR). https://doi.org/10.1109/CVPR.2019.01026

Shao R, Lan X, Yuen PC (2020) Regularized fine-grained meta face anti-spoofing. Proc AAAI Conf Artif Intell 34:11974–11981. https://doi.org/10.1609/aaai.v34i07.6873

Smiatacz M (2012) Liveness measurements using optical flow for biometric person authentication. Metrol Meas Syst 19(2):257–268. https://doi.org/10.2478/v10178-012-0022-y

Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300. https://doi.org/10.1023/A:1018628609742

Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(11):2579–2605

Volpi R, Morerio P, Savarese S, Murino V (2018) Adversarial feature augmentation for unsupervised domain adaptation. IEEE/CVF Conf Comput Vis Pattern Recogn. https://doi.org/10.1109/CVPR.2018.00576

Wang Z, Zhao C, Qin Y, Zhou Q, Lei Z (2018) Exploiting temporal and depth information for multi-frame face anti-spoofing. arXiv:1811.05118

Wang Z, Yu Z, Zhao C, Zhu X, Qin Y, Zhou Q, Zhou F, Lei Z (2020) Deep spatial gradient and temporal depth learning for face anti-spoofing. IEEE/CVF Conf Comput Vis Pattern Recogn (CVPR). https://doi.org/10.1109/CVPR42600.2020.00509

Wen D, Han H, Jain AK (2015) Face spoof detection with image distortion analysis. IEEE Trans Inf Forensics Secur 10(4):746–761. https://doi.org/10.1109/TIFS.2015.2400395

Xu Z, Li S, Deng W (2015) Learning temporal features using lstm-cnn architecture for face anti-spoofing. IAPR Asian Conf Pattern Recogn (ACPR). https://doi.org/10.1109/ACPR.2015.7486482

Yang J, Lei Z, Li SZ (2014) Learn convolutional neural network for face anti-spoofing. arXiv:1408.5601

Yu Z, Li X, Niu X, Shi J, Zhao G (2020) Face anti-spoofing with human material perception. European conference on computer vision. Springer, Berlin, pp 557–575. https://doi.org/10.1007/978-3-030-58571-6_33

Zhang Z, Yan J, Liu S, Lei Z, Yi D, Li SZ (2012) A face antispoofing database with diverse attacks. IAPR Int Conf Biom (ICB). https://doi.org/10.1109/ICB.2012.6199754

Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503. https://doi.org/10.1109/LSP.2016.2603342

Zhao A, Ding M, Liu Z, Xiang T, Niu Y, Guan J, Wen J (2021) Domain-adaptive few-short learning. IEEE Winter Conf Appl Comput Vis (WACV). https://doi.org/10.1109/WACV48630.2021.00143

Zhou L, Luo J, Gao X, Li W, Lei B, Leng J (2021) Selective domain-invariant feature alignment network for face anti-spoofing. IEEE Trans Inf Forensics Secur 16:5352–5365. https://doi.org/10.1109/TIFS.2021.3125603