GENERALIZED FACE ANTI-SPOOFING VIA CROSS-ADVERSARIAL DISENTANGLEMENT WITH MIXING AUGMENTATION

Hanye Huang, Youjun Xiang, Guodong Yang, Lingling Lv, Xianfeng Li, Zichun Weng, Yuli Fu*

School of Electronic and Information Engineering, South China University of Technology
Guangzhou 510641, China

ABSTRACT

Conventional face anti-spoofing methods might be poorly generalized to unseen data distributions. Thus, we improve the generalization of spoof detection from the multi-domain feature disentanglement. Specially, a two-branch convolutional network is proposed to separate spoof-specific features and domain-specific features from face images explicitly. The spoof-specific features are further used for live vs. spoof classification. To minimize correlation among these two features, we present a cross-adversarial training scheme, which requires each branch to act as adversarial supervision for the other branch. To further exploit the subdomains from source data, a mixing augmentation approach is proposed based on mixing domain-specific feature statistics from different instances. It ensures more abundant domain discrepancy and facilitates the disentanglement process. The proposed approach shows promising generalization capacity in several public face anti-spoofing datasets.

Index Terms— Face anti-spoofing, feature disentanglement, adversarial learning, data augmentation

1. INTRODUCTION

Face anti-spoofing plays an important role in defending face recognition systems against presentation attacks such as print, video replay, and 3D mask.

Some solutions for face anti-spoofing have been proposed. The inherent differences in texture [1, 2], image quality [3], or deep semantic features [4, 5, 6] are utilized to distinguish real faces from attacks. These methods have achieved well performance under intra-dataset scenarios but poorly generalize to unseen scenarios because of domain discrepancies. Recently, various domain generalization approaches have been proposed for spoof detection, which learn a shared domain-invariant feature space with multiple source domains as shown in Fig. 1 (a). Shao et al. [7] and Jia et al. [8] aligned different distributions by training the feature generator and the discriminator in an adversarial fashion. Shao et al. [9] proposed a meta-learning framework incorporating domain transfer scenarios. However, these methods were still limited when applied to unseen scenarios with large domain disparities since they may not separate intrinsic patterns between live and spoof faces from irrelevant domain information well [10, 11].

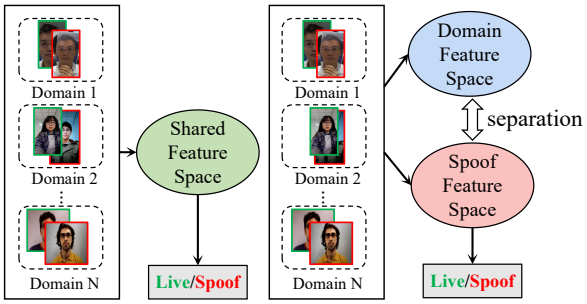


Fig. 1: Methods for multi-domain face anti-spoofing. (a) Previous works tackle the domain generalization problem in the whole feature space. (b) Our work disentangles spoof-specific features and domain-specific features from images, and the live vs. spoof classification could be achieved in a more discriminative feature space.

In this work, we focus on improving the generalization ability of face anti-spoofing via multi-domain disentangled learning. Motivated by [12, 13], we assume that the latent space of face images can be decomposed into spoof-specific and domain-specific features, as shown in Fig. 1 (b). Spoof-specific features contain discriminative information for live vs. spoof classification, while domain-specific features are usually related to illumination, background scenes, recording devices, etc. To achieve this goal, we propose a cross-adversarial disentanglement network. A two-branch convolutional network is constructed to extract spoof-specific and domain-specific features from face images, respectively. To further guarantee that these two features can be split completely, a cross-adversarial training scheme is designed, which requires each branch's classifier to act as adversar-

This work is partially supported by Natural Science Foundation of Guangdong Province (2019A1515010861), Guangzhou Technical Project (Grant: 201902020008), and NSFC (Grant: 61471174).

*Corresponding author.

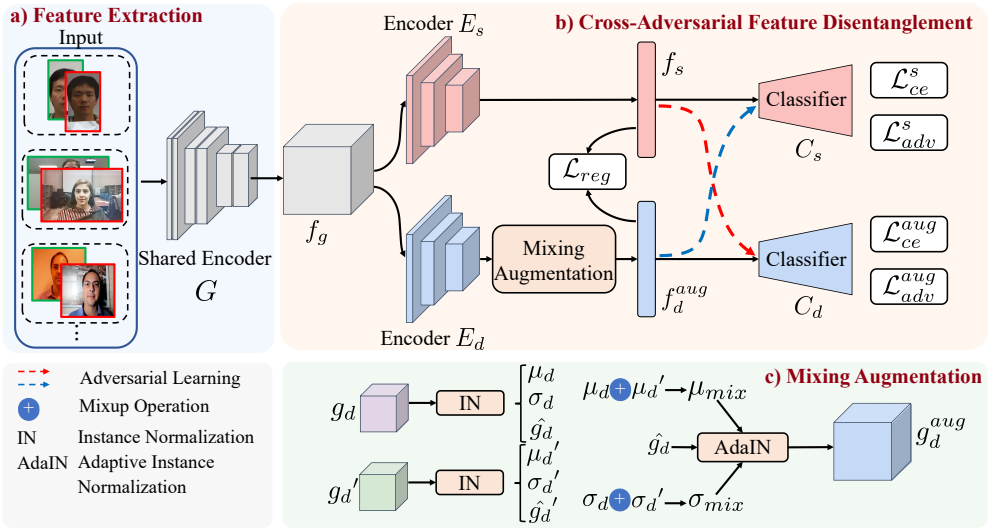


Fig. 2: Detailed overview of the proposed method including cross-adversarial feature disentanglement and mixing augmentation.

ial supervision for the other branch. This design actively minimizes the correlation between two features, resulting in robust spoof-specific representations. In addition, it is unreliable to simply treat a dataset involving various factors as one domain. Otherwise, it could prevent the network from capturing sufficient and intrinsic domain cues. To further exploit the implicit subdomains of the source data, we propose a mixing augmentation method for the domain-specific branch, where novel feature samples are synthesized by mixing the convolutional feature statistics of different instances. The augmented domain feature space leads the network to exclude irrelevant domain attributes, promoting generalization capability.

The main contributions of this work are summarized as follows. (1) We address the face anti-spoofing via multi-domain feature disentanglement, which separates latent representations into spoof-specific features and domain-specific features, allowing more domain-invariant feature representations for live vs. spoof classification. (2) The cross-adversarial training scheme and the mixing augmentation are designed for better disentanglement learning.

2. PROPOSED METHOD

Let (X, Y, D) denote the face images, task labels (live vs. spoof), and domain labels in source domains. Our goal is to improve the generalization capacity of face anti-spoofing via multi-domain disentangled learning.

2.1. Cross-Adversarial Feature Disentanglement

As shown in Fig. 2, a two-branch convolutional network is conducted. Specially, the shared encoder G projects a face image $x \in X$ to a latent representation: $f_g = G(x)$. Two individual encoders E_s and E_d are responsible for encoding spoof-specific feature f_s and domain-specific feature f_d based on f_g, respectively:

$$f_s = E_s(f_g), f_d = E_d(f_g). \tag{1}$$

This structure separates two features from convolving with one another. To actively minimize the correlation between the two types of features, we propose a cross-adversarial learning scheme. The basic structure of the scheme is mirror-symmetrical in design, so we take the spoof-specific feature disentanglement branch as a representative to explain in detail.

First, E_s and the corresponding live vs. spoof classifier C_s are trained to correctly distinguish between real faces and attacks under the supervision of a standard cross-entropy loss:

$$\mathcal{L}^s_{ce} = -\mathbb{E}_{(x,y)\sim(X,Y)} \sum_{k=1}^{K} \mathbb{1}[k = y] log(C_s(f_s)), \tag{2}$$

where K denotes numbers of categories, and $\mathbb{1}[k = y]$ is an indicator function that equals 1 if $k = y$ and 0 else. f_s is expected to contain as little domain information as possible. Intuitively, if the domain classifier C_d can not predict the correct label from f_s, it can be considered that domain-specific factors in f_s have been removed. Therefore, we aim to minimize the negative entropy of the output from C_d with f_s as the input:

$$\mathcal{L}^s_{adv} = -\frac{1}{N} \sum_{x\sim X} log(C_d(f_s)), \tag{3}$$

where N denotes the number of total images. The parameters of C_d are fixed in Eq. (3). This step corresponds to

the red dashed line in Fig. 2. Domain-specific feature disentanglement can be achieved by two-step adversarial training similarly:

$$\mathcal{L}_{ce}^{d} = -\mathbb{E}_{(x,d)\sim(X,D)} \sum_{k=1}^{K_d} \mathbb{1}[k=d] log(C_d(f_d)), \quad (4)$$

$$\mathcal{L}_{adv}^{d} = -\frac{1}{N} \sum_{x\sim X} log(C_s(f_d)), \quad (5)$$

where K_d is the number of domains. The parameters of C_s are fixed in Eq. (5). Eq. (5) corresponds to the blue dashed line in Fig. 2. Note that domain encoder E_d is embedded into Eq. (4) for updating C_d which affects E_s in Eq. (3). E_d is influenced by E_s in the same way. Thus the network works in a mutual boosting way to achieve better feature disentanglement, resulting in generalized spoof-specific features.

2.2. Mixing Augmentation

Recent studies [14, 15] have shown that convolutional feature statistics can represent domain style information to some extent. To ensure more abundant domain discrepancy for training the network, we propose mixing different domain-specific features to generate new samples implicitly.

As shown in Fig. 2, domain-specific feature of an input image before pooling operation can be represented as $g_d \in \mathbb{R}^{C\times H\times W}$. C, H and W denote the size of channel, height and width. The mean μ_d, standard deviation σ_d and normalized feature $\hat{g_d}$ of g_d are computed by instance normalization (IN) [16]. Inspired by [15], we first compute mixed feature statistics to simulate new domain styles through Mixup[17] operation. Given (μ_d, σ_d, $\hat{g_d}$) and ($\mu_d{}'$, $\sigma_d{}'$, $\hat{g_d}{}'$) from two samples $x \in X$ and $x' \in X$, this process can be represented as:

$$\mu_{mix} = \lambda\mu_d + (1-\lambda)\mu_d{}', \quad (6)$$

$$\sigma_{mix} = \lambda\sigma_d + (1-\lambda)\sigma_d{}', \quad (7)$$

where λ is sampled from $\mathrm{Beta}(\alpha, \alpha)$ distribution with α set to 0.1. Then, adaptive instance normalization (AdaIN) [18] is introduced to generate mixed features. Augmented feature g_d^{aug} from g_d can be represented as:

$$g_d^{aug} = \sigma_{mix}\hat{g_d} + \mu_{mix}. \quad (8)$$

Different from [15], we perform mixed feature augmentation on disentangled domain features rather than directly on the entangled convolutional features, which may eliminate beneficial information for spoof detection. We further interpolate the domain label space simultaneously. Therefore, Eq. (4) and Eq. (5) can be reformulated as follows:

$$\begin{aligned}\mathcal{L}_{ce}^{aug} = -\mathbb{E}_{\substack{(x,d)\sim(X,D)\\(x',d')\sim(X,D)}} [\lambda \sum_{k=1}^{K_d} \mathbb{1}[k=d] log(C_d(f_d^{aug})) \\ + (1-\lambda) \sum_{k=1}^{K_d} \mathbb{1}[k=d'] log(C_d(f_d^{aug}))],\end{aligned} \quad (9)$$

$$\mathcal{L}_{adv}^{aug} = -\frac{1}{N} \sum_{x\sim X} log(C_s(f_d^{aug})), \quad (10)$$

where f_d^{aug} denotes domain-specific feature with mixing augmentation.

2.3. Loss Function

To encourage disentanglement, we further impose a soft subspace constraint [19] between two kinds of features as the soft-orthogonal regularization term:

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{x\sim X} \left\| {f_s}^{\mathrm{T}} f_d^{aug} \right\|_F^2. \quad (11)$$

By jointly taking all loss functions into account, the overall objective function is:

$$\mathcal{L} = w_1\mathcal{L}_{ce}^{s} + w_2\mathcal{L}_{adv}^{s} + w_3\mathcal{L}_{ce}^{aug} + w_4\mathcal{L}_{adv}^{aug} + w_5\mathcal{L}_{reg}, \quad (12)$$

where w_1, w_2, w_3, w_4 and w_5 are non-negative parameters to tune the importance of different terms.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets: Following previous works [7, 8], we evaluate the proposed method on four public datasets: Idiap Replay-Attack [20] (**I**), CASIA-FASD [21] (**C**), MSU-MFSD [3] (**M**) and OULU-NPU [22] (**O**). We randomly select three datasets for training and the remaining one for testing. Then we have four testing scenarios: **O&C&M to I**, **O&C&I to M**, **O&M&I to C**, and **I&C&M to O**.

Implementation Details: ResNet-18 [23] is employed as our backbone network. The shared encoder consists of the first three residual convolutional blocks. The encoder and classifier of each branch are implemented by the remaining residual blocks in ResNet-18 and a fully connected layer. We strictly follow the popular evaluation metrics, which contain the Half Total Error Rate (HTER)[24] and the Area Under Curve (AUC).

3.2. Experimental Results

We compare our method with some existing approaches using multiple domains. As shown in Table 1, our method is better than conventional methods, which only focus on fitting source domains. Moreover, our method achieves competitive performance compared to existing domain generalization methods. For instance, our method surpasses PAD-GAN [11] 2.17 percentage points on HTER and 2.55 percentage points on AUC for **O&M&I to C**. The possible reason is that our disentanglement method focuses on separating the discriminative features for spoof detection. In addition, the mixing

Table 1: Comparison results between the proposed method and state-of-the-art methods in four testing scenarios for domain generalization.

Method	References	O&C&M to I		O&C&I to M		O&M&I to C		I&C&M to O	
		HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
IDA [3]	TIFS'15	28.35	78.25	66.67	27.86	55.17	39.05	54.20	44.59
Color Texture [1]	TIFS'16	40.40	62.78	28.09	78.47	30.58	76.89	63.59	32.71
MADDG [7]	CVPR'19	22.19	84.99	17.69	88.06	24.50	84.51	27.98	80.02
SSDG-M [8]	CVPR'20	18.21	**94.61**	16.67	90.47	23.11	85.45	25.17	81.83
PAD-GAN [11]	CVPR'20	20.87	86.72	17.02	90.10	19.68	87.43	25.02	81.47
RFM [9]	AAAI'20	17.30	90.48	13.89	93.98	20.27	88.16	16.45	91.16
D^2AM [14]	AAAI'21	15.43	91.22	12.70	**95.66**	20.98	85.58	15.27	90.87
Ours w/o CA&MA	-	22.62	84.01	16.98	88.71	27.90	79.87	23.62	84.46
Ours w/o CA	-	21.94	83.76	16.61	91.13	26.64	83.19	21.16	86.96
Ours w/o MA	-	17.20	89.26	13.22	92.75	19.21	88.54	15.35	92.43
Ours w/o $\mathcal{L}_{reg}$	-	15.36	90.80	11.71	95.17	17.84	89.83	14.84	92.73
Ours	-	**15.08**	91.92	**11.64**	95.27	**17.51**	**89.98**	**14.27**	**93.04**

Table 2: Comparison to methods with limited domains.

Method	M&I to C		M&I to O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)
IDA[3]	45.16	58.80	54.52	42.17
Color Texture[1]	55.17	46.89	53.31	45.16
MADDG [7]	41.02	64.33	39.35	65.10
SSDG-M [8]	31.89	71.29	36.01	66.88
PAD-GAN [11]	31.67	75.23	34.02	72.65
D^2AM [14]	32.65	72.04	27.70	75.36
Ours w/o CA&MA	44.74	57.86	40.20	58.52
Ours w/o CA	42.58	59.15	35.07	68.73
Ours w/o MA	35.14	68.09	31.73	73.88
Ours w/o $\mathcal{L}_{reg}$	30.64	73.27	28.67	77.49
Ours	**28.75**	**75.78**	**26.28**	**80.46**

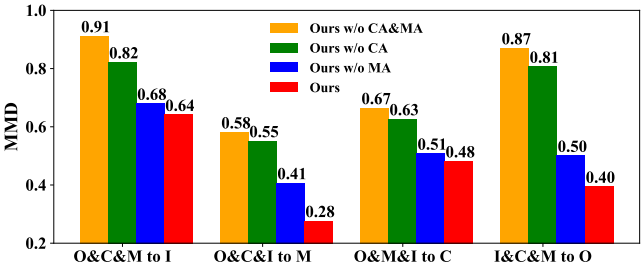
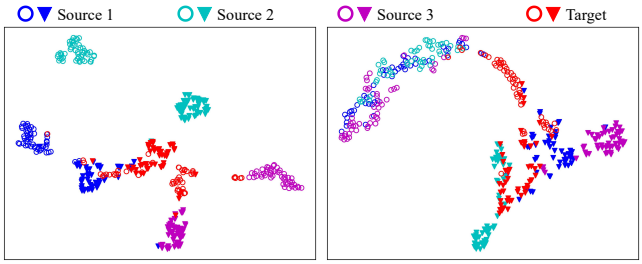


Fig. 3: MMD of f_s space in four testing scenarios.

augmentation helps to regularize the spoof-specific feature space, which facilitates the generalization. As illustrated in Table 2, we also validate the proposed method with limited source domains (i.e., only two source datasets). Our method achieves the best performance in this more challenging case, once again indicating its robustness in unseen domains.

We investigate the impact of each component of our method. The cross-adversarial learning fashion and mixing augmentation are denoted as CA and MA, respectively. Our baseline is designed with $\mathcal{L}_{reg}$ only (i.e., without CA&MA). As shown in Table 1 and Table 2, the most obvious improvements come from our cross-adversarial learning fashion. Since mixing augmentation can expand the domain feature space effectively, it achieves consistent performance improvements. The soft-orthogonal regularization helps to disentangle features further. The best performance can be achieved by combining all of them. In addition, we introduce



(a) f_s (Ours w/o CA&MA) (b) f_s (Ours)

Fig. 4: T-SNE visualizations for the **O&C&M to I** task. Each color represents a domain (red points represent the target domain), circle and triangle points represent real faces and attacks respectively.

maximum mean difference (MMD) [25] to measure the distribution difference between the training and testing datasets. If the MMD value is large, the difference between feature distributions is big. As shown in Fig. 3, the key components of our method enhance the generalization ability of the model.

Visualization: To help understand the effectiveness of disentanglement, we use t-SNE [26] to visualize the features as illustrated in Fig. 4. It can be observed that the entire method contributes to seeking a more domain-invariant feature space to distinguish real faces and attacks.

4. CONCLUSION

In this paper, we focus on improving the generalization capacity of face anti-spoofing via multi-domain feature disentanglement. Specially, we propose a cross-adversarial disentanglement network, which disentangles face images into spoof-specific and domain-specific features explicitly. Mixing augmentation is designed for ensuring more abundant domain discrepancy, which further regularizes the spoof-specific feature space. Experimental results demonstrate the effectiveness of our method against state-of-the-art competitors.

5. REFERENCES

[1] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid, "Face spoofing detection using colour texture analysis," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 8, pp. 1818–1830, 2016.

[2] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid, "Face antispoofing using speeded-up robust features and fisher vector encoding," *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 141–145, 2016.

[3] Di Wen, Hu Han, and Anil K Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

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

[5] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *CVPR*, 2018, pp. 389–398.

[6] Dongmei Peng, Jing Xiao, Rong Zhu, and Ge Gao, "Tsfen: Probing feature selection strategy for face anti-spoofing," in *ICASSP*. IEEE, 2020, pp. 2942–2946.

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

[8] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen, "Single-side domain generalization for face anti-spoofing," in *CVPR*, 2020, pp. 8484–8493.

[9] Rui Shao, Xiangyuan Lan, and Pong C Yuen, "Regularized fine-grained meta face anti-spoofing," in *AAAI*, 2020, vol. 34, pp. 11974–11981.

[10] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu, "On disentangling spoof trace for generic face anti-spoofing," in *ECCV*. Springer, 2020, pp. 406–422.

[11] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *CVPR*, 2020, pp. 6678–6687.

[12] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma, "Face anti-spoofing via disentangled representation learning," in *ECCV*. Springer, 2020, pp. 641–657.

[13] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko, "Domain agnostic learning with disentangled representations," in *ICML*. PMLR, 2019, pp. 5102–5112.

[14] Zhihong Chen, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Feiyue Huang, and Xinyu Jin, "Generalizable representation learning for mixture domain face anti-spoofing," in *AAAI*, 2021, vol. 35, pp. 1132–1139.

[15] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang, "Domain generalization with mixstyle," in *ICLR*, 2021.

[16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[17] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[18] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017, pp. 1501–1510.

[19] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, "Domain separation networks," *Advances in neural information processing systems*, vol. 29, pp. 343–351, 2016.

[20] Ivana Chingovska, André Anjos, and Sébastien Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *BIOSIG*. IEEE, 2012, pp. 1–7.

[21] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li, "A face antispoofing database with diverse attacks," in *ICB*. IEEE, 2012, pp. 26–31.

[22] 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 2017*. IEEE, 2017, pp. 612–618.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] André Anjos and Sébastien Marcel, "Counter-measures to photo attacks in face recognition: a public database and a baseline," in *IJCB*. IEEE, 2011, pp. 1–7.

[25] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[26] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, pp. 2579–2605, 2008.