BLACKBOX FACE RECONSTRUCTION FROM DEEP FACIAL EMBEDDINGS USING A DIFFERENT FACE RECOGNITION MODEL

Hatef Otroshi Shahreza[1,2] *and Sébastien Marcel*[1,3]

[1]Idiap Research Institute, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3]Université de Lausanne (UNIL), Switzerland

ABSTRACT

Face recognition systems generally store features (called embeddings) extracted from each face image during the enrollment stage, and then compare the extracted embeddings with the stored embeddings during the recognition stage. In this paper, we focus on the blackbox face reconstruction from facial embeddings stored in the face recognition database. We use a convolutional neural network (CNN) to reconstruct face images and train our network with a multi-term loss function. In particular, we use a different feature extractor trained for face recognition (which the adversary has the whitebox knowledge of it) to minimize the distance of embeddings extracted from the original and reconstructed face images. We evaluate our method in blackbox attacks against five state-of-the-art face recognition models on the MOBIO and LFW datasets. Our experimental results show that our proposed method outperforms previous face reconstruction methods in the literature. The source code of our experiments is publicly available to facilitate the reproducibility of our work.

***Index Terms*—** blackbox, embedding, face recognition, face reconstruction, template inversion

1. INTRODUCTION

Applications of automatic face recognition systems tend toward ubiquity and in particular, their use in authentication applications is growing rapidly. In such systems, deep neural networks are used to extract some features (a.k.a. "embeddings") from face images. These features are stored in the database of the face recognition system during the enrollment stage, and are later used for comparison during the recognition stage.

Among different types of potential attacks against face recognition systems that have been studied in the literature [1, 2, 3], template inversion attack can jeopardize both the privacy and security of the enrolled users. In the template inversion attack, the adversary gains access to the database of a face recognition system and tries to invert the stored embeddings to reconstruct the underlying face images. The reconstructed face images can reveal privacy-sensitive information about the users, including age, gender, ethnicity, etc. In addition, the adversary can use the reconstructed face image to impersonate the enrolled users and enter the system.

This research is based upon work supported by the H2020 TReSPAsS-ETN Marie Skłodowska-Curie early training network (grant agreement 860813).

Fig. 1: Sample face images from the FFHQ dataset (first row) and their corresponding reconstructed face images from ArcFace embeddings in whitebox (second row) and blackbox (third row) template inversion attacks. The values indicate cosine similarity between embeddings of the original and reconstructed face images. The decision threshold corresponding to FMR $= 10^{-3}$ is 0.37 for ArcFace on the MOBIO dataset.

Generally, methods for reconstruction of face images from embeddings can be categorized into *whitebox* an *blackbox*. In the *whitebox* methods, the adversary is assumed to have complete knowledge of the feature extractor network and its internal parameters. However, in the *blackbox* methods the adversary does not have any knowledge of internal functioning of the feature extractor model and can only use it to extract embeddings for arbitrary images.

In [4], a whitebox face reconstruction method based on gradient-ascent-based optimization (with regularization terms) was proposed. The authors also used the same loss function to train a convolutional neural network (CNN) to reconstruct face images. In [5], also a whitebox face reconstruction method was proposed, where a CNN which was trained with multi-term loss function (including a loss term to minimize the distance between

Table 1: Comparison with related works.

Ref.	Method Basis	Blackbox	Available code
[4]	1) optimization 2) learning	✗	✗
[5]	learning	✗	✓
[6]	learning	✓	✗
[7]	learning	✓	✗
[8]	learning	✓	✓
[9]	learning	✓	✓
[10]	optimization	✓	✓
[11]	optimization	✓	✗
[Ours]	learning	✓	✓

embeddings of the original and reconstructed face images using the feature extractor of the system).

In contrast to [4, 5], in [7, 6] whitebox methods were proposed and were also extended to blackbox attacks. In [7], a multi-layer perceptron (MLP) and CNN were used to estimate landmark coordinates and generate face textures, respectively. Then, a differentiable warping was applied to reconstruct face images. In the whitebox attack, the authors used the warping function and trained the MLP and CNN end-to-end with a multi-term loss function, including a loss term to minimize the distance between embeddings (extracted using the feature extractor of the system) of the original and reconstructed face images. For the blackbox attack, authors trained the MLP and CNN separately, and then reconstructed the face image using the warping function. In [6], a bijection-learning-based approach was used to train a generative adversarial network (GAN) for face reconstruction. While the authors proposed their method based on a whitebox knowledge of feature extractor, they proposed to use distillation of knowledge to train a student network from the face recognition model in the blackbox attack. However, they did not report details on the training of the student network (e.g., the network structure, etc.).

In [8, 10, 9] blackbox (only) methods were proposed to reconstruct face images. In [8], the authors proposed two CNNs based on two new blocks, NBNet-A and NBNet-B, and trained each network structure with two different loss functions to reconstruct face images. In [10], athe authors proposed a greedy random optimization over the latent space of StyleGAN [12] to find a latent vector which synthesizes an image with embedding close to the target embedding. In contrast to [10], in [9], authors trained a MLP to find the latent space of StyleGAN [12] from the embeddings and then generate the face image using pretrained StyleGAN. Similarly, in [11], an optimization-based approach based on the genetic algorithm is proposed to find the latent space of StyleGAN [12] from the embeddings. Table 3 compares our work with previous methods in the literature.

In this paper, we focus on blackbox face reconstruction from facial embeddings. We use the convolutional neural network proposed in [5] to reconstruct face images and train our network with a multi-term loss function. In particular, we propose to use a different feature extractor in our loss function and minimize the embeddings (extracted from this model) of original and reconstructed face images. Fig. 1 shows sample face images from the FFHQ [12] dataset and their whitebox and blackbox reconstructed versions from ArcFace [13] embeddings using our face reconstruction network. We train our face reconstruction network for state-of-the-art face recognition models and evaluate our trained blackbox face reconstruction models on the MOBIO [14] and LFW [15] datasets. The experimental results show that the proposed network outperforms previous blackbox face reconstruction methods in terms of an adversary's success attack rate.

The remainder of the paper is organized as follows. In section 2, we describe our threat model and our proposed blackbox face reconstruction method. Next, we describe our experiments and discuss our results in section 3. Finally, the paper is concluded in section 4.

2. PROPOSED METHOD

2.1. Threat model

We consider the situation where an adversary gains access to the database of embeddings in a face recognition system, and tries to reconstruct the face images from the face embeddings. The adversary is assumed to have the *blackbox* knowledge of the feature extractor of the face recognition system and can use it to extract embeddings from each image. We assume that the adversary has also the *whitebox* knowledge of another feature extractor model, which has been trained for face recognition purpose, and can use this feature extractor model for training the face reconstruction model. The adversary can then use the trained face reconstruction model to invert embeddings and generate face images. The adversary can use the reconstructed face images to inject as a query into the face recognition system.

2.2. Training Data

Let $F_{\text{sys}}(.)$ denote the feature extractor of the target face recognition system that the adversary is assumed to have blackbox knowledge of it. The adversary can use the feature extractor $F_{\text{sys}}(.)$ to generate a training dataset $\mathcal{D} = \{(\mathbf{e}_i, \mathbf{I}_i)\}_{i=1}^{N}$ from a set of face images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$ with N face images, where $\mathbf{e}_i = F_{\text{sys}}(\mathbf{I}_i)$ is the face embedding extracted from image $\mathbf{I}_i$ using $F_{\text{sys}}(.)$. Then, the adversary can use the dataset $\mathcal{D}$ to train a face reconstruction network.

2.3. Face Reconstruction

For our face reconstruction network, we use our network structure proposed in [5], which includes multiple blocks, where each block is composed of 3 cascaded convolutional layers with a skip connection after each deconvolutional layer. We use 6 of these blocks with 512, 256, 128, 64, 32, and 16 filters, and with kernels of sizes 4 and 3 for deconvolution and convolution layers in each block, respectively. In addition, Batch Normalization [16] and a rectified linear unit (ReLU) are used after each deconvolution and convolution operations. Finally, a convolutional layer with a kernel of size 3 and a sigmoid activation function is used to generate the reconstructed face image. Let $\hat{\mathbf{I}}$ denote the reconstructed face

image from embedding $e = F_{sys}(\mathbf{I})$. We train our network with a multi-term loss function including:

- *Mean Squared Error (MSE):* To minimize the pixel-level reconstruction error, we use the square of ℓ_2-norm of the reconstruction error:

$$\mathcal{L}_{\text{MSE}}(\hat{\mathbf{I}},\mathbf{I}) = ||\hat{\mathbf{I}}-\mathbf{I}||_2^2 \tag{1}$$

- *Dissimilarity Structural Index Metric (DSSIM):* To enhance the reconstruction quality in terms of the Similarity Structural Index Metric (SSIM) [17], we use the DSSIM loss term [18] as follows:

$$\mathcal{L}_{\text{DSSIM}}(\hat{\mathbf{I}},\mathbf{I}) = \frac{1-\text{SSIM}(\hat{\mathbf{I}},\mathbf{I})}{2} \tag{2}$$

- *ID loss:* To help the network to reconstruct the face image with similar identity information, we use another feature extractor $F_{\text{loss}}(.)$ trained for face recognition that the adversary has whitebox knowledge of it. Then, we minimize the square of the ℓ_2-norm of the difference between the extracted features from the original and reconstructed face images using $F_{\text{loss}}(.)$:

$$\mathcal{L}_{\text{ID}}(\hat{\mathbf{I}},\mathbf{I}) = ||F_{\text{loss}}(\hat{\mathbf{I}}) - F_{\text{loss}}(\mathbf{I})||_2^2 \tag{3}$$

  It is noteworthy that in whitebox methods, e.g., [5], the same feature extractor of the target system is used for ID loss, but we propose to use a different feature extractor for the blackbox attack (i.e., $F_{\text{loss}} \neq F_{\text{sys}}$).

For our total loss, we use a linear summation of the aforementioned loss term:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{\text{DSSIM}} + \beta\mathcal{L}_{\text{ID}}, \tag{4}$$

where α and β are hyperparamters, and are experimentally found perform the best for $\alpha = 0.1$ and $\beta = 0.005$.

3. EXPERIMENTS

3.1. Experimental Setup

We use state-of-the-art face reconstruction models including ArcFace [13] and ElasticFace [19] as well as three different FR models with state-of-the-art backbones from FaceX-Zoo [20], including AttentionNet [21], HRNet [22], and Swin [23]. Table 2 reports recognition performance of these models on the MOBIO and LFW datasets. We use ArcFace and ElasticFace as F_{loss} and evaluate the performance of our method in blackbox attack against other models.

To train our face reconstruction networks, we use the FFHQ [12] dataset, which includes 70,000 face images (90% train and 10% validation). We also evaluate our models on the MOBIO [14] and LFW [15] datasets. The MOBIO dataset includes face images captured by mobile devices from 152 people and the LFW dataset includes 13,233 images of 5,749 people. We use the face recognition models and build face recognition systems on the MOBIO and LFW datasets. Then, we use our reconstruction model trained on FFHQ to invert enrolled embeddings and reconstruct face images. We inject the reconstructed face image as a query to the system to evaluate the performance of face reconstruction in terms of an adversary's Success Attack Rate (SAR) in entering the system when the system is configured at False Match Rate (FMR) of 10^{-3}.

Table 2: Recognition performance of face recognition models in terms of true match rate (TMR) at false match rates (FMRs) of 10^{-2} and 10^{-3} evaluated on the MOBIO and LFW datasets.

model	MOBIO		LFW	
	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}
ArcFace	100.00	99.98	97.60	96.40
ElasticFace	100.00	100.00	96.87	94.70
AttentionNet	99.71	97.73	84.27	72.77
HRNet	98.98	98.23	89.30	78.43
Swin	99.75	98.98	91.70	87.83

We use the Bob toolbox [24] and PyTorch package in our implementations. To train our face reconstruction networks, we use the Adam [25] optimizer with the initial learning rate of 10^{-3}, and we decrease the learning rate every 10 epochs, by a factor of 0.5. The source code of our experiments is publicly available to help reproduce our results[1].

3.2. Comparison with previous methods

As described in section 3.1, we use ArcFace and ElasticFace as F_{loss} and train our model to reconstruct face images from embeddings extracted by other face recognition models. Table 3 compares the performance of our method with previous blackbox methods[2] in terms of SAR. As this table shows, our method outperforms previous blackbox face reconstruction methods in the literature. Also, comparing the results for ArcFace and ElasticFace as F_{loss} in our loss function, the networks trained with ArcFace achieve better performance than the networks trained with ElasticFace. This might be due to superior recognition performance of ArcFace compared to ElasticFace as reported in Table 2

3.3. Discussion

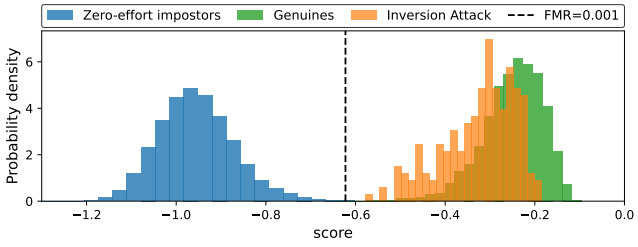
Fig. 1 illustrates sample face images from the validation set of the FFHQ dataset and their reconstructed face images from ArcFace embeddings in the whitebox and blackbox attacks. In the whitebox attack, we used the same feature extractor (i.e., ArcFace) as F_{loss} (similar to [5]) and in the blackbox attack, we used a different feature extractor (i.e., ElasticFace) in our ID loss. Fig. 2 also shows the histogram of scores between ArcFace embeddings extracted from genuine and zero-effort impostor pairs, as well as the original and reconstructed face images in the whitebox and blackbox attacks evaluated on the MOBIO dataset. As this figure shows, the score distribution for reconstructed face images of

[1]Source code: https://gitlab.idiap.ch/bob/bob.paper.icip2023_blackbox_face_reconstruction

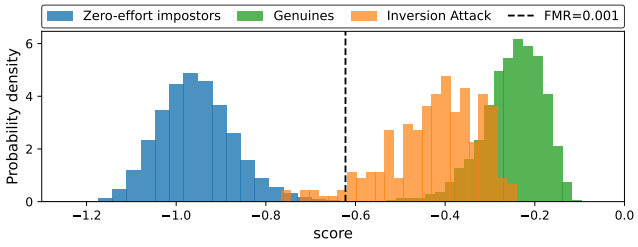
[2]As indicated in Table 1, other previous blackbox methods such as [7, 6, 11] do not have available source code.

Table 3: Performance comparison with previous blackbox face reconstruction method in terms of adversary's success attack rate (SAR) at system configured at false match rate (FMR) of 10^{-3}. In each case, the best two values are emboldened.

method	MOBIO					LFW				
	ArcFace	ElasticFace	AttentionNet	HRNet	Swin	ArcFace	ElasticFace	AttentionNet	HRNet	Swin
NBNetA-M [8]	0	2.38	0	0	0	4.32	10.90	1.24	1.60	3.82
NBNetA-P [8]	4.76	16.19	0.48	0	7.14	16.83	26.98	0.66	1.44	9.70
NBNetB-M [8]	1.90	3.80	3.33	7.14	8.57	10.98	21.44	3.22	4.47	11.23
NBNetB-P [8]	15.24	**43.81**	31.90	26.67	44.29	40.26	**58.16**	16.29	18.42	40.76
Dong *et al.* [9]	3.33	8.10	10.48	6.67	3.33	13.21	12.61	3.90	4.07	12.38
Vendrow and Vendrow [10]	**29.05**	**43.81**	27.14	26.67	45.24	**57.70**	53.03	21.12	18.85	46.84
[Ours] (F_{loss}= ElasticFace)	**95.71**	-	**89.05**	**93.81**	**98.10**	**90.67**	-	**58.85**	**65.55**	**80.20**
[Ours] (F_{loss}= ArcFace)	-	**98.10**	**92.38**	**97.62**	**99.52**	-	**92.27**	**62.20**	**68.99**	**82.21**



(a) whitebox



(b) blackbox

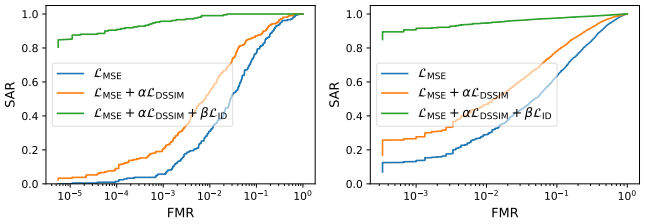
Fig. 2: Histogram of scores (negative cosine distance) in the (a) whitebox and (b) blackbox attacks against ArcFace evaluated on the MOBIO dataset.

whitebox attack is slightly closer to genuine distribution than that of blackbox attack, which is expected as we have the whitebox knowledge of the system and use the same feature extractor in the training. Meanwhile, the score distribution for reconstructed face images of blackbox attack is still closer to the score distribution of genuine pairs than the distribution of zero-effort impostor pairs.

To evaluate the effect of each term in our loss function, as an ablation study, we train our network with different loss terms in Eq. 4. Fig. 3 compares the reconstruction performance of models trained with different loss functions in terms of SAR for different values of the system's FMR. According to these results, the ID loss and the DSSIM loss terms improve the reconstruction performance. In particular, the ID loss (using another feature extractor method) significantly enhances the SAR in our method.

4. CONCLUSION

In this paper, we proposed a blackbox method to reconstruct face images from facial embeddings using a CNN-based structure. We used a multi-term loss function and in particular proposed to use another feature extractor trained for face recognition and minimize the embeddings extracted from the original and reconstructed face images. We evaluated our face reconstruction method for five state-of-the-art face recognition models on the MOBIO and LFW datasets. The experimental results show that our method outperforms previous blackbox face reconstruction methods in terms of the adversary's success attack rate in entering the face recognition system. Furthermore, our ablation study shows that applying ID loss using a different feature extractor can significantly improve the success attack rate.



(a) MOBIO (b) LFW

Fig. 3: Ablation study on the effect of each term in our loss function on reconstruction performance using ArcFace embeddings evaluated on (a) MOBIO and (b) LFW datasets.

5. REFERENCES

[1] Battista Biggio, Paolo Russu, Luca Didaci, Fabio Roli, et al., "Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 31–41, 2015.

[2] Javier Galbally, Chris McCool, Julian Fierrez, Sebastien Marcel, and Javier Ortega-Garcia, "On the vulnerability of face verification systems to hill-climbing attacks," *Pattern Recognition*, vol. 43, no. 3, pp. 1027–1038, 2010.

[3] Sébastien Marcel, Julian Fierrez, and Nicholas Evans, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, Springer, 2023.

[4] Andrey Zhmoginov and Mark Sandler, "Inverting face embeddings with convolutional neural networks," *arXiv preprint arXiv:1606.04189*, 2016.

[5] Hatef Otroshi Shahreza, Vedrana Krivokuća Hahn, and Sébastien Marcel, "Face reconstruction from deep facial embeddings using a convolutional neural network," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1211–1215.

[6] Chi Nhan Duong, Thanh-Dat Truong, Khoa Luu, Kha Gia Quach, Hung Bui, and Kaushik Roy, "Vec2face: Unveil human faces from their blackbox features in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6132–6141.

[7] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman, "Synthesizing normalized faces from facial identity features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3703–3712.

[8] Guangcan Mai, Kai Cao, Pong C Yuen, and Anil K Jain, "On the reconstruction of face images from deep face templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 5, pp. 1188–1202, 2018.

[9] Xingbo Dong, Zhe Jin, Zhenhua Guo, and Andrew Beng Jin Teoh, "Towards generating high definition face images from deep templates," in *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2021, pp. 1–11.

[10] Edward Vendrow and Joshua Vendrow, "Realistic face reconstruction from deep embeddings," in *Proceedings of NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

[11] Xingbo Dong, Zhihui Miao, Lan Ma, Jiajun Shen, Zhe Jin, Zhenhua Guo, and Andrew Beng Jin Teoh, "Reconstruct face from features using gan generator as a distribution constraint," *arXiv preprint arXiv:2206.04295*, 2022.

[12] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[13] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] Chris McCool, Roy Wallace, Mitchell McLaren, Laurent El Shafey, and Sébastien Marcel, "Session variability modelling for face authentication," *IET Biometrics*, vol. 2, no. 3, pp. 117–129, Sept. 2013.

[15] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[16] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning (ICML)*, Lille, France, Jul. 2015, pp. 448–456.

[17] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] Sahar Sadrizadeh, Hatef Otroshi-Shahreza, and Farokh Marvasti, "Impulsive noise removal via a blind cnn enhanced by an iterative post-processing," *Signal Processing*, vol. 192, pp. 108378, 2022.

[19] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1578–1587.

[20] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei, "Facex-zoo: A pytorch toolbox for face recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[21] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3156–3164.

[22] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al., "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021, pp. 10012–10022.

[24] A. Anjos, M. Günther, T. de Freitas Pereira, P. Korshunov, A. Mohammadi, and S. Marcel, "Continuously reproducing toolchains in pattern recognition and machine learning experiments," in *Proceedings of the International Conference on Machine Learning (ICML)*, Aug. 2017.

[25] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, California., USA, May 2015.