Thermal to Visible Facial Image Translation Using Generative Adversarial Networks

Zhongling Wang, *Student Member, IEEE*, Zhenzhong Chen, *Senior Member, IEEE*, and Feng Wu, *Fellow, IEEE*

***Abstract*—Thermal cameras can capture images invariant to illumination conditions. However, thermal facial images are difficult to be recognized by human examiners. In this letter, an end-to-end framework, which consists of a generative network and a detector network, is proposed to translate thermal facial images into visible ones. The generative network aims at generating visible images given the thermal ones. The detector can locate important facial landmarks on visible faces and help the generative network to generate more realistic images that are easier to be recognized. As demonstrated in the experiments, the faces generated by our method have good visual quality and maintain identity preserving features.**

***Index Terms*—Face, generative adversarial network (GAN), image translation, infrared, thermal.**

I. Introduction

THERMAL cameras can work under different illumination conditions. However, thermal facial images are hard to be recognized by human examiners because of the large modality gap between the visible and the thermal images. Fig. 1 shows a thermal image x and a visible image y of the same individual. As we can see, it is challenging to identify the person in the thermal image due to the large modality gap and the lack of details in the thermal image.

In this letter, we provide a new solution to generate visible images, which maintain identity preserving features while at the same time have good visual quality. We incorporate two networks into our framework: the generative network and the detector network. The generative network aims at translating thermal images to visible ones. The detector network can extract facial landmarks (see S' and S'' in Fig. 1) from the generated visible images and can propagate its loss back to the generative network. With the help of the gradient information of the detector network, the generative network will generate more realistic visible images. Our model outperforms other generative models in terms of visual and face recognition evaluations. In addition, a dataset containing 1584 aligned thermal and visible facial images is established.

Manuscript received April 5, 2018; revised May 24, 2018; accepted May 28, 2018. Date of publication June 8, 2018; date of current version June 28, 2018. This work was supported by National Key R&D Program of China under contract 2017YFB1002202. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wei Li. *(Corresponding author: Zhenzhong Chen.)*

Z. Wang is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: zhlwang@whu.edu.cn).

Z. Chen is with the School of Computer Science, Wuhan University, Wuhan 430072, China, and also with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China (e-mail: zzchen@whu.edu.cn).

F. Wu is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: fengwu@ustc.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2018.2845692

II. Related Work

There are much fewer works investigating thermal–visible facial image translation than near infrared-visible facial image translation [1], [2] due to the larger modality gap between the thermal and visible images [3]. Conventional thermal–visible translation methods [4], [5] can generate a grayscale visible facial image given a thermal input. Li *et al.* [4] proposed a learning-based framework synthesizing the visible face by exploiting the local linearity in the image spatial domain and the image manifolds. Dou *et al.* [5] learned the relationship between feature spaces using locally linear regression. More recently, generative adversarial networks (GANs) [6] based general-purpose image translation methods are also worth noticing. Many of them can be applied to thermal–visible translation. Based on GANs, Pix2pix [7] further incorporated the L_1 distance between the generated and the ground truth images into their model. This made it possible to generate images maintaining more low-frequency information. Moreover, CycleGAN [8] introduced cycle consistency into their model, which can learn a bidirectional translation between two image distributions.

Similar to our work, some works [9], [10] have utilized GANs to handle thermal–visible facial images translation. Based on Pix2pix, Zhang *et al.* [9] further incorporated an explicit closed-set face recognition loss, which can preserve more identity information. Based on GANs, Zhang *et al.* [10] used an identity loss and a perceptual loss to translate thermal facial images to visible ones. The authors in [9] and [10] mainly focused on generating images preserving facial features suitable for recognition. However, high face recognition accuracies of those generated images do not always mean they also have good visual qualities.

III. Proposed Method

A. Framework Overview

Our model mainly consists of two components: the generative network and the detector network. The generative network alone can translate thermal images to visible ones, but in a coarse manner. Our generative network is based on CycleGAN [8]. It learns the bidirectional translation between thermal and visible images in an unsupervised manner using unpaired training images. The two translations form a translation cycle and can benefit each other during training. We choose CycleGAN instead of Pix2pix as our generative network because in practice Pix2pix is more likely to suffer from overfitting when a rather small dataset

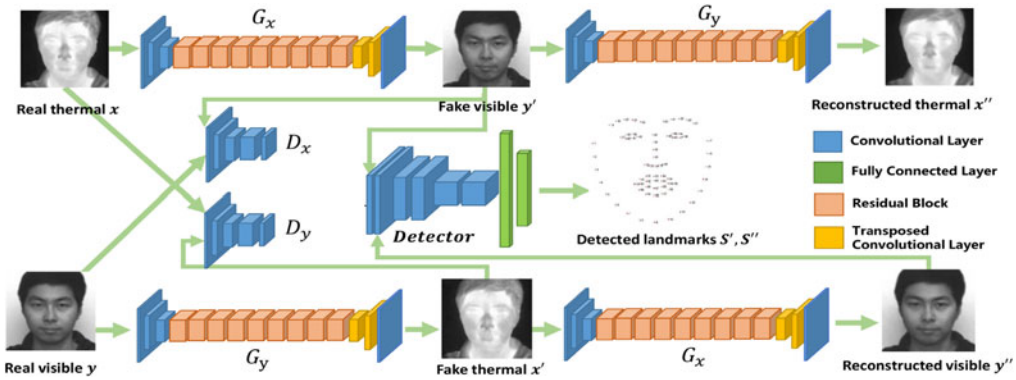

Fig. 1. Framework of our proposed method.

is used for training. Pix2pix is also more sensitive to image misalignment due to the L1 loss it used. The detector network can extract face shape features, which are constituted by important landmarks of visible faces (see S' and S'' in Fig. 1). Working as constraints of the generative network, these shape features can help the generative network to generate visible images of better visual quality and with more realistic identity preserving features. Different from the face recognition loss and the perceptual loss used in [9] and [10], which were intended for improving face recognition performance, the loss of the detector network can not only help to preserve identity features but also help to generate photo-realistic images.

Our overall framework is illustrated in Fig. 1, which is constituted by the generative network and the detector network. The generative network is composed of two generators (G_x, G_y) and two discriminators (D_x, D_y). The detector network is an CNN that can locate 68 facial landmarks on a visible face. Detailed descriptions about each part in Fig. 1 are provided in the following two sections. Our objective function L_{all} incorporates a generic image translation loss and a face specific loss, which can be formulated as

$$L_{\text{all}} = L_{\text{CycleGAN}} + \gamma L_{\text{shape}} \tag{1}$$

where γ is a parameter balancing the two terms. L_{CycleGAN} is a generic loss that measures the distance between two image distributions, which is directly introduced from CycleGAN [8]. However, since it is difficult for us to figure out the implicit judgment criteria of the discriminators due to the nature of GANs, and the loss of details in thermal images increases the difficulty of the translation, in practice the L_{CycleGAN} loss alone may fail to generate images of satisfying quality. Therefore, a face shape constraint L_{shape} is introduced in our model to help the generative network to improve its performance.

B. Generative Network

Facial images are sensitive to distortions and artifacts. So generated images with photo-realistic appearance are desired. Inspired by CycleGAN [8], which is an unsupervised generative model, we introduce it into our framework as the generative network. The basic idea of CycleGAN is to explore the cycle consistency between two different image distributions. In our case, the two distributions are the thermal facial image distribution and the visible facial image distribution. The generative network is constituted by G_x, G_y, D_x, and D_y. G_x and G_y are two generators. G_x translates grayscale thermal images x to grayscale visible ones y', which can be formulated as $y' = G_x(x)$. G_y translates visible images y to thermal ones x': $x' = G_y(y)$. D_x and D_y are two discriminators. D_x can evaluate the qualities of visible images, while D_y can evaluate thermal images. Following [8], we can formulate the objectives of the two GANs as

$$L_{\text{GANs}} = \frac{1}{2}\left[\left((D_x(y)-1)^2 + D_x(y')^2\right) + (D_y(x)-1)^2 + D_y(x')^2\right].$$

Then another two translations $x'' = G_y(y')$ and $y'' = G_x(x')$ are performed. The cycle consistency loss can be formulated as $L_{\text{cyc}_x} = \|x'' - x\|_1$ and $L_{\text{cyc}_y} = \|y'' - y\|_1$. Following CycleGAN [8], the overall objective of the generative network is

$$L_{\text{CycleGAN}} = L_{\text{GANs}} + \lambda(L_{\text{cyc}_x} + L_{\text{cyc}_y}) \tag{2}$$

where λ is a parameter. The L_{CycleGAN} loss alone can perform a basic transformation. But in order to maintain more photo-realistic and identity preserving features in the generated images, an additional loss is needed.

C. Detector Network

The generative network is a generic model, which is capable of translating images of any domain to another. However, as a well known fact that GANs are difficult to train [11], it is challenging to get satisfying results on the thermal-visible translation task using the generative network alone. The implicit judgment criteria of the discriminators in the generative network make it difficult for us to figure out what kind of image will the discriminators consider as "realistic."

Although thermal cameras are robust to different illumination conditions, they still have some disadvantages that may increase the difficulty of the translation. The most conspicuous difficulty is that thermal images contain less detail and texture than visible images due to the emissive nature of thermal imagery and the low resolution of thermal sensors. In addition, the overheating of face regions will make the thermal images partially overexpose and lose lots of detail.

In practice, directly employing the generative network alone (CycleGAN [8]) or some other generic models like Pix2pix [7] to our problem usually results in visible faces with noticeable distortions and artifacts (see Section IV). So we introduce a facial landmark detector into our model to improve the

performance by locating key landmarks of human faces. These landmarks can depict fundamental structures and identity preserving features of human faces [12], [13]. We follow Deng *et al.*'s [14] method to build our detector. The loss of the detector, which is also called the shape loss, can be formulated as

$$L_{\text{shape}} = \frac{1}{68}\left[\left(S - S'\right)^2 + \left(S - S''\right)^2\right] \tag{3}$$

where S' and S'' represent sets of 68 detected landmarks of the generated images y' and y'', respectively. S is the set of 68 ground truth landmarks detected using the Openface toolkit [15] (see Section IV). Each landmark is an x, y coordinate. In each iteration, landmarks of the visible faces generated by the generative network are extracted. Then the shape loss L_{shape} is calculated and propagated back to the generative network. In this way, the detector can help the generative network to learn the latent structure of human faces, and, hence generate more realistic images. Although the shape loss looks similar to the perceptual losses used by [10] and [16]–[18], which are formed by features extracted from intermediate layers of networks, they are different. The meanings behind these intermediate features have not been understood thoroughly while our landmark feature has a clear meaning: the deviations of positions of landmarks. In addition, Li *et al.* [19] used a similar parser detector to handle visible face completion. This parser detector can segment a visible face into different regions. However, the pretrained parser detector remains fixed during training, while the landmark detector in our model is trained together with other parts of the model. This allows the landmark detector to be more adaptive to the visible images generated by the generative network in the early stage of training.

IV. Experiments

Since we need to calculate L_{shape}, ground truth landmarks of the visible and thermal faces are required. Landmarks of visible faces can be easily obtained using many off-the-shelf detectors [15]. However, obtaining landmarks of thermal faces is not a trivial task. An alternative solution is to utilize aligned visible and thermal image pairs: landmarks of the visible and thermal images are the same. But aligned datasets are rare. Therefore, we establish a new datase[1] containing 792 aligned thermal and visible facial image pairs of 33 individuals. A FLIR AX5 thermal camera is used to obtain the thermal images. A rectangular 6×4 checkerboard with a width of 2 m and a height of 1.5 m containing 24 rectangles is placed 2 m in front of these individuals. Individuals being photographed are asked to face toward each of the 24 rectangles. For each rectangle, a visible image and a thermal image are taken simultaneously. The original sizes of the visible and thermal images are 992×794 and 320×256, respectively. We use OpenFace [15] to detect visible faces and locate 68 facial landmarks on each visible face. The detected face regions of both spectra that serve as the ground truth visible y and thermal x images are converted to grayscale and resized to 128×128. The dataset is randomly divided into two sets: the training set containing 552 image pairs of 23 subjects and the testing set containing 240 image pairs of the other 10 subjects. Images in the training set are used only for training and those in the testing set are used only for testing.

[1]Available at iip.whu.edu.cn/projects/IR2Vis_dataset.html

In practice, after tuning hyperparameters of CycleGAN [8] and Pix2pix [7] on our dataset, we found the default hyperparameters produce very competitive results. So in the following evaluations, default hyperparameters and architectures are used in CycleGAN [8] and Pix2pix [7]. Since the generative network in our model is derived from CycleGAN, to be fair, all the shared hyperparameters and architectures in our model and those in CycleGAN [8] are set to the same: we employ the generators containing nine residual blocks and the 70×70 PatchGAN [7] discriminators. The default $\lambda = 10$ is used in (2). We use $\gamma = 0.0001$ in (1). In each iteration during training, we use Adam optimizer [20] to update each network only once following the order: G_x, G_y, D_x, D_y, and Detector.

In the following two sections, our model is compared against other two state-of-the-art image translation models: CycleGAN [8] and Pix2pix [7]. According to (1) and (2), we can write the loss of our model as $L_{\text{GANs}} + L_{\text{cyc}} + L_{\text{shape}}$, where $L_{\text{cyc}} = L_{\text{cyc}_x} + L_{\text{cyc}_y}$. In order to demonstrate the function of each part of our model, we also provide ablation experiments with different loss function combinations given in the following:

1) $L_{\text{GANs}} + L_{\text{cyc}}$ (CycleGAN [8]).
2) $L_{\text{cyc}} + L_{\text{shape}}$.
3) L_{cyc}.
4) $L_{\text{GANs}} + L_{\text{shape}}$.
5) L_{GANs}.

Data augmentation (random crop and random horizontal flip) is used only in the training phases of all methods. Except for the supervised Pix2pix, the image pairs after augmentation may no longer be aligned.

A. Visual Performance

We show the images translated from the testing thermal images in Fig. 2. Partially enlarged regions are shown in Fig. 3. In Fig. 3, first row, the eyes in column (c) and (d) contain noticeable artifacts. In the second row, the eyebrow in column (c) and the eye in column (d) are distorted. In the third row, column (c) contains a big dark spot on the forehead, which is conspicuous in Fig. 2. However, benefiting from the detector network, the locations and shapes of facial features are more accurate in our images.

Generated visible images are also compared against ground truth visible images using structural similarity (SSIM) [21], multiscale (MS-SSIM) [22], peak signal to noise ratio (PSNR), and mean squared error (MSE). The average scores of the testing set are depicted in Table I. Our method outperforms other approaches in all the four evaluations. The SSIM and MS-SSIM performances demonstrate the superiority of our method in preserving structural information that constitutes the textures of faces.

Facial landmarks can depict the shape of facial features. In order to quantitatively demonstrate that our model can maintain more accurate facial features, we evaluate the landmark accuracy of the generated visible images. To be fair, the detector we use in this experiment is the OpenFace toolkit [15]. MSE of the 68 detected landmarks of each face are depicted in Table I. According to the result, the capability of our model preserving accurate locations of facial features is very competitive. We also show the result of landmark detection of our model in Fig. 4. The landmarks of the generated images are close to those of the

Fig. 2. Different methods for thermal–visible facial image translation. (a) Thermal images. (b) Ours ($L_{\text{GANs}} + L_{\text{cyc}} + L_{\text{shape}}$). (c) CycleGAN [8] ($L_{\text{GANs}} + L_{\text{cyc}}$). (d) Pix2pix [7]. (e) $L_{\text{cyc}} + L_{\text{shape}}$. (f) L_{cyc}. (g) $L_{\text{GANs}} + L_{\text{shape}}$. (h) L_{GANs}. (i) Ground truth. Partially enlarged images of columns (a), (b), (c), (d), and (i) are shown in Fig. 3.

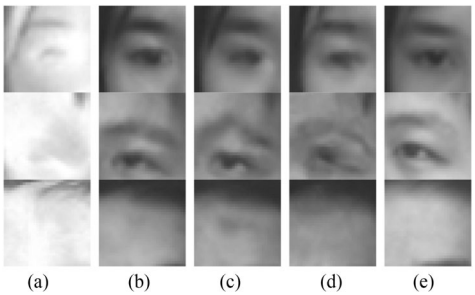

Fig. 3. Enlarged regions of the images in Fig. 2. (a) Thermal images. (b) Ours. (c) CycleGAN [8]. (d) Pix2pix [7]. (e) Ground truth.

TABLE I
QUANTITATIVE EVALUATIONS

Method	SSIM	MSSSIM	PSNR	MSE	Landmark MSE
Ours	0.785	0.828	21.41	724.1	330.7
CycleGAN [8]	0.752	0.802	19.49	985.8	385.1
Pix2pix [7]	0.742	0.798	19.07	949.5	468.4
$L_{cyc} + L_{shape}$	0.298	0.192	11.67	4551.6	-
L_{cyc}	0.177	0.124	9.03	8284.9	-
$L_{GANs}+L_{shape}$	0.757	0.808	19.81	866.8	297.6
L_{GANs}	0.754	0.802	19.63	920.9	378.7

Fig. 4. Detected landmarks (red) and the ground truth landmarks (blue) of the generated images. (Best viewed in color.)

ground truth images, which further affirm the usefulness of the detector network.

B. Face Recognition Performance

In this section, the ability of our method generating images preserving accurate identity information is demonstrated. Facenet,[2] which is pretrained on public available visible datasets, is employed for face recognition without any modification. In our testing dataset, 240 ground truth visible images of 10 people serve as the gallery pool, while 240 generated visible images serve as the probe images. To be fair, each probe image's correspondence ground truth visible face is removed from the gallery pool. For each probe image, its correspondence gallery pool contains 239 images. We randomly choose one image from the gallery pool for each subject to form our gallery. Our face recognition evaluation has three steps. First, features of the generated probe and gallery images are obtained by directly forwarding them through the Facenet model without realignment. Second, for each probe image, the Euclidean distance between its features and the features of all gallery images are calculated. Third, we take the shortest distance as the predicted match and determine whether it is correct or not. This face recognition experiment is repeated 100 times since the gallery is randomly chosen. The mean recognition accuracy of all probe images in the testing set is reported in Table II. We also provide two experiments as baselines using the same Facenet recognition model as follows.

1) Low light visible. We reduce the brightness (multiply 0.2 to each pixel) of the visible gallery images. Then these darkened visible gallery images serve both as probe images and gallery images.
2) Raw thermal. Thermal images in the testing set without translation serve as probe images and the gallery is composed of randomly chosen visible images.

TABLE II
AVERAGE RECOGNITION ACCURACY (NOTATIONS ARE ADOPTED FROM THE CUMULATIVE MATCH CURVE)

Accuracy	Rank1	Rank3	Rank5
Ours	0.916	0.993	0.999
CycleGAN[8]	0.889	0.988	0.998
Pix2pix[7]	0.774	0.958	0.995
$L_{cyc} + L_{shape}$	0.300	0.645	0.886
L_{cyc}	0.218	0.449	0.682
$L_{GANs} + L_{shape}$	0.779	0.971	0.997
L_{GANs}	0.788	0.966	0.995
Low light visible	0.104	0.298	0.486
Raw thermal	0.081	0.278	0.476

V. CONCLUSION

We propose an approach to translate thermal facial images into visible ones. Our framework consists of two parts: the generative network and the detector network. The generative network learns a coarse mapping from the thermal images to the visible images. The detector network locates key landmarks on visible faces and helps the optimization of the generative network. As shown in our experiments, our method outperforms other recent state-of-the-art methods.

[2] Available at https://github.com/davidsandberg/facenet, schroff2015facenet

REFERENCES

[1] N. Wang, J. Li, D. Tao, X. Li, and X. Gao, "Heterogeneous image transformation," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 77–84, 2013.

[2] J. Lezama, Q. Qiu, and G. Sapiro, "Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6807–6816.

[3] M. S. Sarfraz and R. Stiefelhagen, "Deep perceptual mapping for cross-modal face recognition," *Int. J. Comput. Vis.*, vol. 122, no. 3, pp. 426–438, May 2017.

[4] J. Li, P. Hao, C. Zhang, and M. Dou, "Hallucinating faces from thermal infrared images," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 465–468.

[5] M. Dou, C. Zhang, P. Hao, and J. Li, "Converting thermal infrared face images into normal gray-level images," in *Proc. 8th Asian Conf. Comput. Vis.-vol. Part II*, 2007, pp. 722–732.

[6] L. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[7] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.

[8] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.

[9] T. Zhang, A. Wiliem, S. Yang, and B. C. Lovell, "TV-GAN: Generative adversarial network based thermal to visible face recognition," *in Proc. 11th IAPR Int. Conf. Biometrics*, 2018.

[10] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu, "Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2017, pp. 100–107.

[11] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[12] J. Shi, A. Samal, and D. Marx, "How effective are landmarks and their geometry for face recognition?" *Comput. Vis. Image Understanding*, vol. 102, no. 2, pp. 117–133, 2006.

[13] Z. Zhang, L. Wang, Q. Zhu, S. Chen, and Y. Chen, "Pose-invariant face recognition using facial landmarks and weber local descriptor," *Knowl.-Based Syst.*, vol. 84, pp. 78–88, 2015.

[14] Z. Deng, K. Li, Q. Zhao, Y. Zhang, and H. Chen, "Effective face landmark localization via single deep network," in *Proc. Chin. Conf. Biometric Recognit.*, 2017, pp. 68–76.

[15] T. Baltrusaitis, P. Robinson, and L. P. Morency, "Openface: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2016, pp. 1–10.

[16] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and superresolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[17] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 105–114.

[18] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," arXiv: 1701.05957, 2017.

[19] Y. Li, S. Liu, J. Yang, and M. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, vol. 1, no. 3, pp. 1–13.

[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. Learn. Represent.*, 2015.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[22] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, vol. 2, pp. 1398–1402.

[23] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.