2018 International Conference on Biometrics

TV-GAN: Generative Adversarial Network Based Thermal to Visible Face Recognition

Teng Zhang[*], Arnold Wiliem[*], Siqi Yang and Brian C. Lovell
Security and Surveillance Group, The University of Queensland
{patrick.zhang, a.wiliem, siqi.yang}@uq.edu.au, lovell@itee.uq.edu.au

Abstract

This work tackles the face recognition task on images captured using thermal camera sensors which can operate in the non-light environment. While it can greatly increase the scope and benefits of the current security surveillance systems, performing such a task using thermal images is a challenging problem compared to face recognition task in the Visible Light Domain (VLD). This is partly due to the significantly smaller amount of thermal imagery data collected compared to the VLD data. Unfortunately, direct application of the existing very strong face recognition models trained using VLD data into the thermal imagery data will not produce a satisfactory performance. This is due to the existence of the domain gap between the thermal and VLD images. To this end, we propose a Thermal-to-Visible Generative Adversarial Network (TV-GAN) that is able to transform thermal face images into their corresponding VLD images whilst maintaining identity information which is sufficient enough for the existing VLD face recognition models to perform recognition. Some examples are presented in Figure 1. Unlike the previous methods, our proposed TV-GAN uses an explicit closed-set face recognition loss to regularize the discriminator network training. This information will then be conveyed into the generator network in the form of gradient loss. In the experiment, we show that by using this additional explicit regularization for the discriminator network, the TV-GAN is able to preserve more identity information when translating a thermal image of a person which is not seen before by the TV-GAN.

1. Introduction

Face recognition is one of the most important tasks in a smart video surveillance systems and it has been extensively studied in the visible light domain (VLD). Recently, the existing deep neural network based VLD face recognition systems have achieved impressive performance [23, 28, 33] due to the advent of extremely large face datasets [17, 34]. With these great strides, it is imperative to extend the existing VLD based face recognition systems into other less studied domains such as near-infrared imaging (low-light) and thermal imaging (no-light).

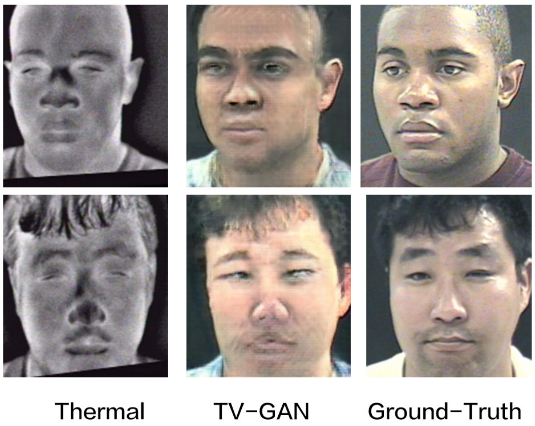

Figure 1. Example results. The first column is input thermal faces and the second column are the generated faces. Note that these people have never been seen in the model training phase. Compared with the ground-truth from the last column, we can see that the proposed method can recover considerable detail.

The difference between these three domains is illustrated in Figure 2. Whilst, the thermal face has lost most of the texture and edge information, the near-infrared face has significant similarity to the VLD face images. As such, performing the face recognition task in the thermal image domain is significantly more challenging than in the near-infrared image domain.

Unfortunately, the above-mentioned successes in the VLD domain could not be easily replicated in the thermal domain due to the relatively small amount of training data available and the domain gap between thermal and visible light. As shown in our experiment, directly applying the VLD based face recognition systems will not achieve satisfactory performance. Being able to work with pre-existing

[*] Denote equal contribution

0-7695-6403-8/18/31.00 ©2018 IEEE
DOI 10.1109/ICB2018.2018.00035

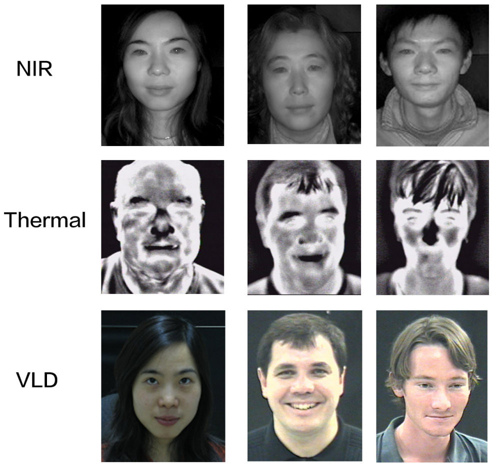

Figure 2. Randomly selected face image samples in three different domains.

face recognition systems is extremely important. In fact, it requires enormous effort to train a very accurate face recognition system.

To this end, our strategy is to utilize image transformation techniques to the thermal query images. Once transformed, an existing pre-trained VLD deep neural network face recognition model can be directly employed as a black box. So, instead of using the thermal face, the generated "fake" VLD face will be fed into the VLD face recognition model. With this pre-processing strategy, we can achieve much better recognition results without changing or retraining the VLD model. The framework of our proposed method is sketched in Figure 3. Whilst we agree that a generative process is a much harder problem to solve, solving the problem can address the enormous effort to collect data in the thermal domain and retrain face recognition system. With the advent of the Generative Adversarial Network (GAN) method, the community has made significant progress in solving the generative problem [12] [31].

To achieve good recognition performance, we train a generator network using the Generative Adversarial Network (GAN). Here, we use an adversary network, denoted a discriminator network, aiming to discriminate between real or fake/generated images, to train the generator. In order to preserve the identity information of the person in the transformed image, we add an identity loss which is based on the closed-set face recognition task into the discriminator network. Unlike the previous work in [36] that uses perception loss [13] to indirectly measure the identity loss, our identity loss is more explicit. Note that the perception loss requires training another classifier network such as VGG network [30] and thus makes the training significantly more expensive.

Contributions - Our contributions can be listed as follows. The main contribution is that we propose a novel Thermal-to-Visible Generative Adversarial Network (TV-GAN) that is able to preserve sufficient identity information when transforming thermal images into their corresponding VLD images. We draw our inspiration from the recent Pix2Pix [12] and DR-GAN [31]. Our key idea is to train our discriminator to perform both the binary classification of fake or real and to perform a closed-set face recognition task. We validate our method in challenging experiment scenarios and still achieve better performance than the other recent methods. TV-GAN reduces the domain shift between the VLD and the thermal domain and thereby improves the existing VLD face recognition systems without changing anything. Our source code is publicly available online at https://github.com/uqtzhan2/tv-gan.

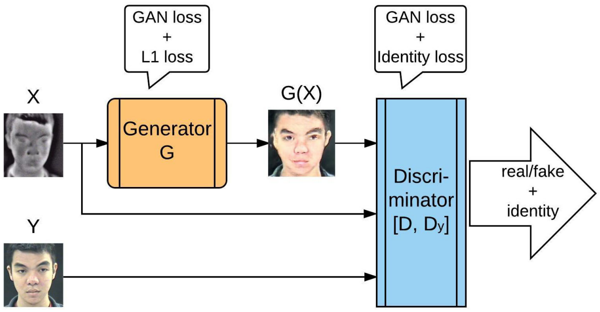

Figure 3. The proposed TV-GAN structure: Our discriminator not only provides the discrimination of fake and real, but also performs a closed-set face recognition task. The generator network G aims to produce images (called "fake" images) that can "fool" the discriminator. Note that our multi-task discriminator is composed of two networks with shared weights $[\mathrm{D}, \mathrm{D}_{\boldsymbol{y}}]$, where D discriminates whether $\boldsymbol{Y}$ is real or not and $\mathrm{D}_{\boldsymbol{y}} : \boldsymbol{X} \times \boldsymbol{Y} \mapsto \{0,1\}^{N+1}$ performs the closed-set face recognition task.

This paper is continued as follows: Section 2 will introduce the related works and the proposed approach is described in Section 3. Experimental results are provided and discussed in Section 4. Finally, we conclude this paper in Section 5.

2. Related Works

As discussed above, face recognition is an extensively studied area in the visible light domain. Recently, many researchers have attempted to address the same problem in other domains such as near-infrared domain [35] [14] [18], polarmetric thermal domain [10] [25] [29] and conventional thermal domain [27] [2] [16]. A few selected related works are discussed below in two categories: non-deep learning learning and deep learning.

Non-deep learning: Before the deep learning era, a few works were proposed to train traditional classifiers based on the feature representations which aim to reduce domain gap [3] [19] [16] [2] [9]. In [19], a simple method that hallucinates the visible face from the thermal face was proposed by exploiting the local linearity in not only the image spatial domain but also the image manifolds. Choi *et al.* [3] presented a partial least square (PLS) based regression framework to model the large modality difference in the PLS latent space. Similarly, Hu *et al.* [9] used a discriminant PLS based approach by specifically building discriminant PLS gallery models for each subject using thermal cross examples. Klare *et al.* [16] describe a face as a vector of kernel similarities to a set of prototypes. In [2], a cascaded subspace learning scheme which consists of whitening transformation, factor analysis and common discriminant analysis is proposed. Despite good performance exhibited by these works, they require the face recognition systems to be retrained. Furthermore, most works only consider frontal aligned face region images without challenging lights, expressions and poses.

Deep learning: Recent methods utilizing deep neural networks have been proposed [27] [18] [12] [36]. For instance, Sarfraz *et al.* [27] used a neural network to learn a reverse mapping, from VLD to mid-wave and low-wave infrared, so that a thermal face image could be matched to its VLD counterpart. This strategy has the disadvantage of having to apply the mapping to each VLD image in the dataset. Different from this work, we propose to use a conditional generative adversarial network that only requires to transform the query thermal image into the VLD image. Similar with traditional machine learning, deep learning based image transforming methods have been investigated [18] [12] [36]. For instance, Lezama *et al.* [18] proposed to use a patch-based transform Convolutional Neural Network (CNN) to hallucinate a visible face from a near-infrared face. The work from [12], known as Pix2Pix, can transform images from domain A to domain B. The idea is to train a Conditional Generative Adversarial Network (CGAN) that can “fool” domain B classifier using a processed image from domain A. However, this work did not ensure that the identity information was preserved during the transformation.

Perhaps the most similar work to us is the work from Zhang *et al.* [36] which also employed the GAN to generate faces from the thermal images. However, there are significant differences between our framework and theirs. Firstly, their work calculates the identity loss indirectly using high-level semantic features extracted from a classification network. This requires an additional network during training. Secondly, we consider a much more challenging scenario wherein the visible light images are in color instead of grey scale and the faces have various pose and occlusion with eyeglasses. In addition, the images used in our evaluation are not perfectly aligned and included head, neck, and part of the chest.

3. Proposed Framework

In this section, a brief description of the Generative Adversarial Networks will be presented. Then, the proposed TV-GAN is elucidated.

3.1. Generative Adversarial Networks

Generative Adversarial Network (GAN) is first introduced in [4]. It comprises two models with competing tasks. The first model, the generator model G, aims to generate an image which resembles a real image. The aim for the second model, the discriminator D is to separate between the fake images from the generator and the real images. Generally, both models are represented by deep neural networks.

Since its first introduction, GAN has been extended into various applications. For instance, Mirza *et al.* [21] propose the Conditional GAN of which the generator learns the data distribution condition upon an input. Radford *et al.* [24] proposed a class of GAN that can stabilize the training. Recently GAN has also been extended for generating images from text description [37], generating style and structure of natural indoor scene images [32] and translating an image from one domain to the other [12].

In this work, we use the conditional GAN framework that allows transforming an image from one domain to the other domain. The GAN architecture used in this work can be briefly described as follows. Let $\mathrm{D} : \boldsymbol{X} \times \boldsymbol{Y} \mapsto \{fake, real\}$, $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{w \times h}$ be the discriminator function and $\mathrm{G} : \boldsymbol{X} \mapsto \boldsymbol{Y}$ be the generator. Note that, unlike the original GAN description in [4] that uses a Gaussian random vector $\boldsymbol{z}$ as the input for the generator function, we follow the GAN architecture described in [12] which uses dropout to maintain the sample diversity. In addition, according to [12], the generator G will generate a better image when it is trained with a discriminator D admitting two inputs: the original image $\boldsymbol{X}$, and the transformed image $\boldsymbol{Y}$. The transformed image $\boldsymbol{Y}$ can be either from the ground truth when $\boldsymbol{Y} \sim P_{data}$, or generated by using $\mathrm{G}(\boldsymbol{X})$. The architecture is then trained using the following objective:

$$\mathcal{L}_{cGAN}(\mathrm{G}, \mathrm{D}) = \mathbb{E}_{\boldsymbol{X} \sim P_{data}(\boldsymbol{X})}\left[\log \mathrm{D}(\boldsymbol{X}, \boldsymbol{Y})\right] + \mathbb{E}_{\boldsymbol{X} \sim P_{data}}\left[\log 1 - \mathrm{D}(\boldsymbol{X}, \mathrm{G}(\boldsymbol{X})\right] \quad (1)$$

where $\mathcal{L}_{cGAN}$ is the conditional GAN loss function.

3.2. TV-GAN: Thermal-to-Visible GAN

The goal of Thermal-to-Visible GAN (TV-GAN) is to train a generator G that will transform a thermal image $\boldsymbol{X}$ into its corresponding visible image $\boldsymbol{Y}$ of which the visible image $\boldsymbol{Y}$ still carries sufficient identity information for face recognition task. To this end, we base our method on Pix2Pix [12]. Pix2Pix is able to transform an image from one domain to the other domain. Unlike the CGAN, Pix2Pix is able to generate sharper images due to its additional loss function that explicitly penalizes the deviation of the generated image $\mathrm{G}(\boldsymbol{X})$ from the ground truth $\boldsymbol{Y}\sim P_{data}$. Unfortunately, Pix2Pix does not have explicit regularization that helps to preserve the personal identity. Recent work in [31] that proposes Disentangled-Representation GAN (DR-GAN) shows that it is possible to improve feature discrimination for the face recognition task by explicitly adding an identity loss function to the discriminator training loss function. The efficacy of using identity loss has also been shown in the GAN-based Visible Face Synthesis (GAN-VFS) [36]. The difference is that GAN-VFS calculates the loss indirectly by using the perceptual loss [13] which uses high-level semantic features extracted from a classification network such as the VGG network [30].

Different from GAN-VFS, we use the more explicit identity loss function, similar to the DR-GAN which is aimed to learn disentangled feature representation solely from VLD images. More specifically, we define our discriminator as a multi-task discriminator that does not only provide the discrimination fake or real but also performs a closed-set face recognition task. We note that although we train the discriminator to perform a closed-set face recognition task, the aim here is to use the gradient information from the discriminator to train the generator so it can generate visible images with sufficient identity information of the person for the recognition task. Later in the experiment part, we will show that this approach is still effective for performing the face recognition tasks where the query person has not been seen by the TV-GAN. Let $\boldsymbol{y} \in \{0,1\}^{N+1}$ be a one-hot-encoding $(N+1)$-dimensional identity vector wherein if the p-th element of vector $\boldsymbol{y}_i$ is 1, then the image $\boldsymbol{X}_i$ belongs to the p-th person; N is the number of subjects in the training set and we reserve additional dimension for the generated images. This way, the discriminator only learns the identity information from the real images. Our multi-task discriminator is composed of two networks with shared weights $[\mathrm{D}, \mathrm{D}_{\boldsymbol{y}}]$, where D discriminates whether $\boldsymbol{Y}$ is real or not and $\mathrm{D}_{\boldsymbol{y}} : \boldsymbol{X} \times \boldsymbol{Y} \mapsto \{0,1\}^{N+1}$ performs the closed-set face recognition task.

The proposed TV-GAN training loss is defined as follows:

$$\begin{aligned}&\mathcal{L}_{TV-GAN}(\mathrm{G},\mathrm{D},\mathrm{D}_{\boldsymbol{y}}) = \\ &\mathcal{L}_{cGAN}(\mathrm{G},\mathrm{D}) + \lambda_1\mathcal{L}_{\ell_1}(\mathrm{G}) + \lambda_2\mathcal{L}_{id}(\mathrm{G},\mathrm{D}_{\boldsymbol{y}}),\end{aligned} \quad (2)$$

where $\mathcal{L}_{cGAN}$ is defined in (1), $\mathcal{L}_{\ell_1}$ is the additional loss function between the generated image and the corresponding visible image ground-truth from Pix2Pix. This loss is required to make sure that the generated exemplars have the distribution close to the visible image ground-truth distribution. $\mathcal{L}_{\ell_1}$ is defined as:

$$\mathcal{L}_{\ell_1}(\mathrm{G}) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}\sim P_{data}}\left[\|\boldsymbol{Y} - \mathrm{G}(\boldsymbol{X})\|_1\right]. \quad (3)$$

The identity loss function is defined as follows:

$$\mathcal{L}_{id}(\mathrm{G},\mathrm{D}_{\boldsymbol{y}}) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}\sim P_{data}}\left[\log(\mathrm{D}_{\boldsymbol{y}}(\boldsymbol{X},\boldsymbol{Y})\right] \quad (4)$$

As for the network architecture, we adopt both the Pix2Pix's generator and discriminator networks without modification. In particular, the generator network uses the U-Net network [26], which is an encoder-decoder with skip connection between mirrored layers in the encoder and decoder stacks.

4. Experiments and Results

In this section, we first describe the implementation details and baselines. Then, the dataset and evaluation protocol will be presented. Finally, we provide analysis based on the performance of various methods.

4.1. Implementation

All evaluations were done by using an NVidia K40c GPU with the tensorflow framework. In addition, Adam optimizer [15] was used with a batch size of 1. Following [12], all networks were trained from the scratch with learning rate of 0.0002, $\beta_1 = 0.5$. As for TV-GAN, we trained the network with 65 epochs. The hyperparameters λ_1 and λ_2 in [4] were set to 100 to make the loss terms in the same scale.

As for the VLD face recognition network, we used pre-trained MatConvNet VGG-based model from [23] without any fine-tuning. We call this VGG-face. The query of the VGG-face is a transformed image $\boldsymbol{Y} = f(\boldsymbol{X})$, where $\boldsymbol{X}$ is the image in the thermal domain and $\boldsymbol{Y}$ is the transformed image. All the images in the gallery $\mathcal{G} = \{\boldsymbol{X}_m\}_{m=1}^{M}$ are VLD images.

4.2. Baselines

In the evaluation, three baselines were used:

Plain Thermal - No transformation was applied on this baseline. In other words, for this baseline, the function f only does the identity mapping, $\boldsymbol{Y} = \boldsymbol{X}$. Thus, essentially, this baseline will indicate the effect of the domain gap between thermal and visible light to the face recognition models such as VGG-face trained solely under VLD images.

Patch based method - It has been shown in [18] that it is possible to learn transformation function $f(\cdot)$ for Near Infrared Domain to VLD by using CNN based with encoder-decoder architecture. We apply this for thermal-to-visible conversion. More specifically, a set of paired image patches were first extracted and then the CNN was trained based on these patches. In this experiment, the patch size 25×25 pixels was used. For a fair comparison, we did not apply the post-processing method blending images from both domains to obtain better performance. As for the CNN architecture, we opted to use a more recent CNN architecture called RedNet [20] which shares similarities to the U-Net [26]. The difference is that RedNet has a skip connection the same as ResNet [5] whereas U-Net has a skip connection the same as DenseNet [11]. From our empirical evaluation (not shown here) both skip connection types gave similar performance. We used RedNet20 which has 20 layers and trained with 108 epochs. The difference between the Patch-based Transform and TV-GAN generator function G is how the networks are trained. The Patch-based Transform method used the mean-squared error loss, whereas TV-GAN used the Generative Adversarial loss.

Pix2Pix [12] - As mentioned in the previous section that Pix2Pix does not have the identity loss regularization in the training whilst the proposed TV-GAN has this regularization. Since Pix2Pix is a GAN-based method, the transformation function f is its generator function, $f = \text{G}$. We trained Pix2Pix using 85 epochs.

4.3. Dataset and evaluation protocol

We used the IRIS dataset [22] for the evaluation in which all images were captured by FLIR SC6700 (spectral range 3um – 7 um). There were 29 subjects with 4,228 pairs of thermal/visible images. As the subjects have various poses, we excluded repeated angles, extreme poses, expressions, and illumination for our experiments. In total, there were 695 pairs of roughly aligned thermal/visible images ($695 \times 2 = 1,390$ images in total) of 29 subjects used in our experiments for all methods.

In the experiment, we needed to measure how much identity information is preserved by the transformation function f for each method. To this end, the recognition accuracy metric was used. We measured this based upon the accuracy of the VGG-face in recognizing the query $\boldsymbol{Y} = f(\boldsymbol{X})$ with the gallery $\mathcal{G}$. More specifically, we used VGG-face to extract features from $\boldsymbol{Y}$. Then, nearest neighbor search with cosine distance metric was used.

There were two protocols used in the evaluation. Protocol A: each subject in the gallery only had one frontal face; and Protocol B: each person in the gallery had four images covering several pose angles. The experiment was repeated multiple times and the average performance was reported. In addition, we also reported rank one, three, five and seven performance. For each repeat, the dataset was randomly divided into 8 subjects for testing and 21 subjects for training the transformation function f (approximately 500 images). The eight subjects in the testing set only had the thermal faces and the corresponding visible faces were excluded from the experiment. We then enrolled the visible faces of all the 29 subjects into the VGG-Face. Note that we did not use the visible faces of the eight subjects that correspond to the thermal faces used by the TV-GAN to generate the visible images. This protocol ensured that no subject images were in training and testing sets during training the transformation function f.

The gallery used by the VGG-face always contained 29 subjects with a different number of images (*i.e.* one image each subject for protocol A and four images each subject for protocol B). The visible face gallery images have never been seen by the pre-trained face recognition model, which is downloaded from the VGG website. This is because the model was trained on the CelebFaces dataset. The test set for each split was used as the queries for the VGG-face. The transformation functions for TV-GAN and Pix2Pix were trained on each split training set. For a fair comparison, we did not perform any data augmentation such as horizontal flipping, rotation and cropping for the gallery images used by the VGG-face. The data augmentation was only applied for training the transformation function f.

4.4. Results

Table 1. Average recognition accuracy (in %) for the setting where only one frontal face visible image for each person is available (Protocol A).

Accuracy	Rank 1	Rank 3	Rank 5	Rank 7
Plain Thermal	3.5	12.2	21.9	27.2
Patch based	8.9	18.9	26.9	36
Pix2Pix	12.1	28.8	39.2	47
TV-GAN no L1	10.71	28.9	39.14	47.21
TV-GAN	**13.9**	**33**	**46.8**	**53.4**

Table 2. Average recognition accuracy (in %) for the setting where four face visible images for each person are available (Protocol B)

Accuracy	Rank 1	Rank 3	Rank 5	Rank 7
Plain Thermal	4.9	15.6	21.45	26.5
Patch based	14.6	23.5	30.1	35.7
Pix2Pix	16	30.7	37.3	44.9
TV-GAN no L1	19.34	34.21	40.84	45.89
TV-GAN	**19.9**	**35.8**	**45.6**	**50.9**

The results for protocol A are presented in Table 1 and Figure 4. The results for protocol B are reported in Table 2 and Figure 5. Additionally, we provide the ROC curve in Figure 6.

As we can see from these results the proposed TV-GAN outperforms all the methods. Also, it is noteworthy to mention that all methods outperform the plain thermal method by a large margin. This suggests the importance of reducing the domain gap existed between thermal and VLD images if one plans to utilize face recognition systems trained solely in the VLD for recognizing faces in the thermal domain. Note that the overall performance on this dataset is quite low for all methods. This is due to the fact that it is a very difficult dataset with non-aligned face pairs including changing lights, extreme poses, occlusion and expression.

The improvement from patch-based method to GAN-based methods such as Pix2pix and the proposed TV-GAN suggests the efficacy of the GAN loss for this application. Upon a closer look, the generators trained using GAN loss could generate much better images compared to the patch-based method. We conjecture that this might be caused by the following factors: (1) The thermal images may not carry sufficient information compared to the near infrared images; (2) the dataset contains large pose angles and (3) the paired thermal and visible images in the training data are not well aligned. As stated in [18], the patch-based method requires well-aligned pairs of images.

Finally, as the proposed TV-GAN outperforms Pix2Pix, this indicates that the closed-set face recognition regularization is more effective to train a generator that can preserve the personal identity in the generated images. By comparing TV-GAN no L1 with TV-GAN results, $\mathcal{L}_{\ell_1}$ loss is shown to play an important role in making sure that the generated exemplars have the distribution close to the visible image ground-truth distribution.

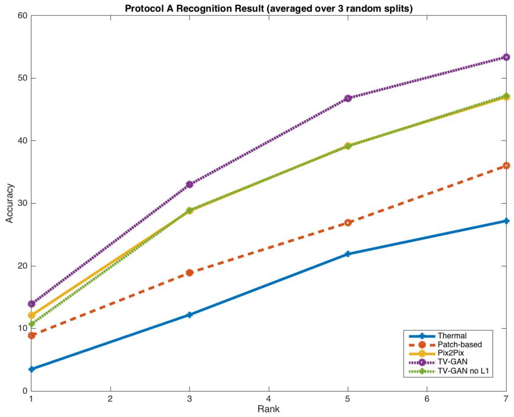

Figure 4. The gallery setting: one visible light image per person (protocol A). We run 3 splits and calculate the average results.

Visualization - Here we present some visualization of all the methods. As we can see from Figure 7, the patch-based method generates the worst visible images. Both Pix2Pix and TV-GAN are able to generate reasonable visualization. However, it is noteworthy to mention that both TV-GAN and Pix2Pix do not always generate the correct attributes. For instance, the race of subject C and the age of subject D are incorrectly inferred. This could be due to the absence of regularization guiding the GAN training in preserving these face attributes.

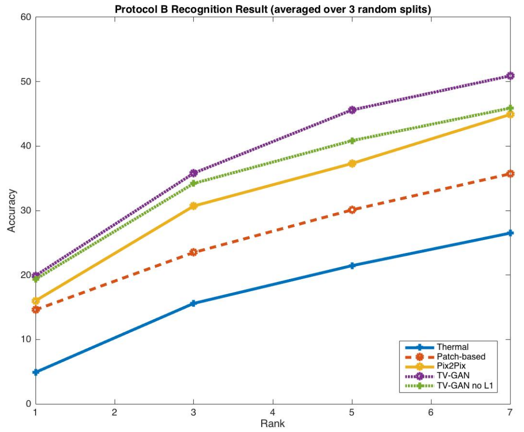

Figure 5. The gallery setting: four visible light image per person (protocol B). We run 3 splits and calculate the average results.

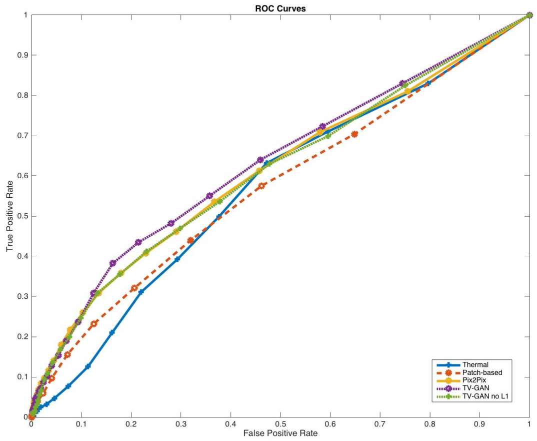

Figure 6. ROC curves for all methods.

Table 3. We deliberately exclude all people with eye glasses from the training set but put them all in the testing set to undermine the generator network of GAN.

Accuracy	Rank 1	Rank 3	Rank 5	Rank 7
Plain Thermal	**12.57**	**36.65**	**46.07**	**54.97**
Pix2Pix	7.85	22.51	36.13	41.88
TV-GAN	12.04	24.08	31.41	37.7

Role of training data - Despite its good performance, the proposed TV-GAN method is trained using a relatively small amount of images. Here we show the importance of having sufficient training data. To emulate a data starvation scenario we deliberately exclude all people with eyeglasses from the training data. The results are reported in Table 3. The results suggest that both Pix2Pix and TV-GAN could not achieve good performance. Surprisingly, the plain thermal method outperforms both GAN methods. This could be due to the fact that the subjects in the testing set wear eyeglasses in both query and gallery images. This information may have been picked up by the features extracted from the VGG-face.

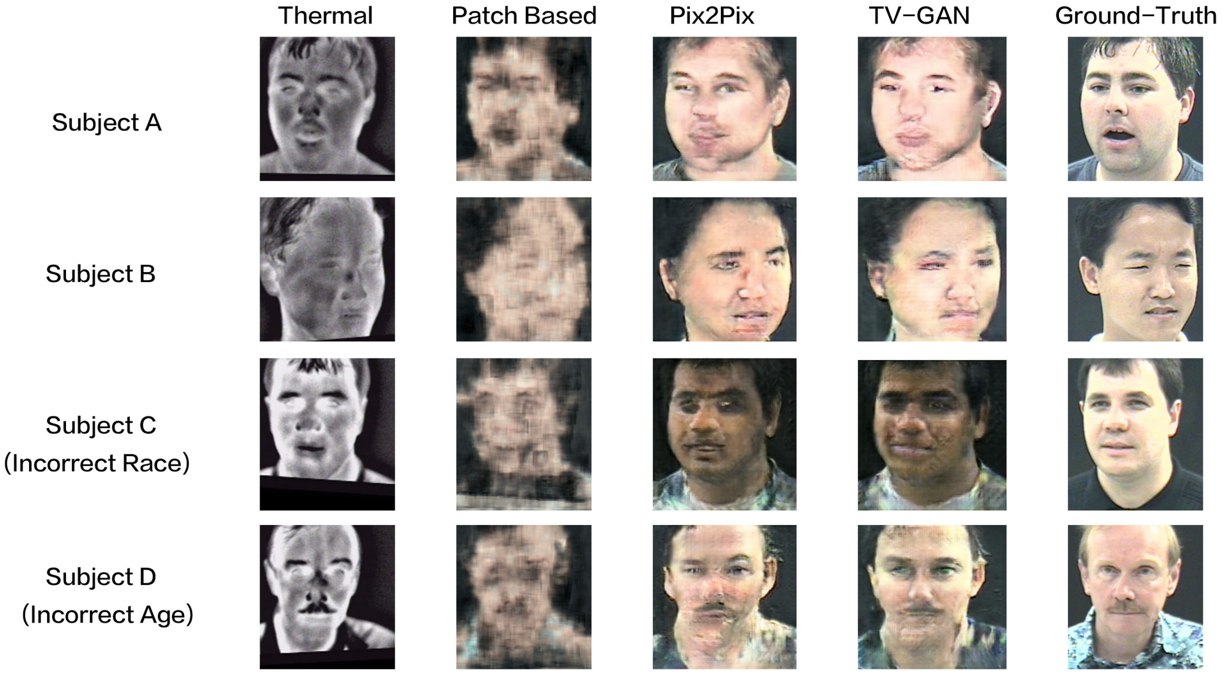

Figure 7. Visualization of selected results. It is noteworthy to mention that both TV-GAN and Pix2Pix do not always generate the correct attributes such as race for subject C and age for subject D.

5. Conclusion

Thermal to visible face recognition/verification task is a challenging problem as little data is available, and applying a face recognition model exclusively trained on visible light domain images produces a non-satisfactory performance. In this paper, we developed a GAN-based method that can apply pre-trained VLD deep learning face recognition models with no further fine tuning. More specifically, a generator network was trained using the Generative Adversarial Network framework which has two networks such as generator and discriminator networks trained against each other. The discriminator guided the generator via its gradient information. Our key insight was that by using a closed-set face recognition task loss inserted into the discriminator, it allowed the generator to learn the transformation function that preserved sufficient identity information for the VLD face recognition system. Despite the existence of challenges such as occlusions, expressions, high pose, different skin tone and limited training data, in our experiment, we showed that our TV-GAN method outperformed the other methods.

Our proposed TV-GAN method is still far from perfect as it still did not ensure the correct transfer of the other face attributes such as race and age in the transformed images. This will be investigated in the future. In addition, we are interested in applying the proposed TV-GAN method for applications in medical imaging such as [1, 6, 7, 8]

6. Acknowledgements

This work has been funded by the Australian Research Council (ARC) Linkage Projects Grant LP160101797. Arnold Wiliem is funded by the Advance Queensland Early-Career Research Fellowship. We also would like to acknowledge Danny Smith for helpful discussions.

References

[1] J. Carvajal, D. F. Smith, K. Zhao, A. Wiliem, P. Finucane, P. Hobson, A. Jennings, R. McDougall, and B. C. Lovell. An early experience toward developing computer aided diagnosis for gram-stained smears images. In *CVPRW*, 2017.

[2] C. Chen and A. Ross. Matching thermal to visible face images using hidden factor analysis in a cascaded subspace learning framework. *Pattern Recognition Letters*, 72(Supplement C):25 – 32, 2016.

[3] J. Choi, S. Hu, S. S. Young, and L. S. Davis. Thermal to visible face recognition. *Proc.SPIE*, 8371:8371 – 8371 – 10, 2012.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *NIPS*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[6] P. Hobson, B. C. Lovell, G. Percannella, A. Saggese, M. Vento, and A. Wiliem. Computer aided diagnosis for anti-

nuclear antibodies hep-2 images: progress and challenges. *Pattern Recognition Letters*, (82):3–11, 2016.

[7] P. Hobson, B. C. Lovell, G. Percannella, A. Saggese, M. Vento, and A. Wiliem. Hep-2 staining pattern recognition at cell and specimen levels: Datasets, algorithms and results. *Pattern Recognition Letters*, (82):12–22, 2016.

[8] P. Hobson, B. C. Lovell, G. Percannella, M. Vento, and A. Wiliem. Classifying anti-nuclear antibodies HEp-2 images: A benchmarking platform. In *ICPR*, 2014.

[9] S. Hu, J. Choi, A. L. Chan, and W. R. Schwartz. Thermal-to-visible face recognition using partial least squares. *Journal of the Optical Society of America A*, 32(3):431–442, 2015.

[10] S. Hu, N. J. Short, B. S. Riggan, C. Gordon, K. P. Gurton, M. Thielke, P. Gurram, and A. L. Chan. A polarimetric thermal database for face recognition research. In *CVPRW*, 2016.

[11] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[13] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[14] F. Juefei-Xu, D. K. Pal, and M. Savvides. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *CVPRW*, 2015.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

[16] B. F. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1410–1422, June 2013.

[17] G. B. H. E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[18] 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 *CVPR*, 2017.

[19] J. Li, P. Hao, C. Zhang, and M. Dou. Hallucinating faces from thermal infrared images. In *ICIP*, 2008.

[20] X.-J. Mao, C. Shen, and Y.-B. Yang. Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections. In *NIPS*, 2016.

[21] M. Mirza and S. Osindero. Conditional generative adversarial nets. In *ArXiv*, 2014.

[22] U. of Tennessee. *EEE OTCBVS WS Series Bench; DOE University Research Program in Robotics under grant DOE-DE-FG02-86NE37968*. 2012.

[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[24] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[25] B. S. Riggan, N. J. Short, S. Hu, and H. Kwon. Estimation of visible spectrum faces from polarimetric thermal faces. In *BTAS*, 2016.

[26] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[27] M. S. Sarfraz and R. Stiefelhagen. Deep perceptual mapping for cross-modal face recognition. *IJCV*, 122(3):426–438, 2017.

[28] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[29] N. Short, S. Hu, P. Gurram, K. Gurton, and A. Chan. Improving cross-modal face recognition using polarimetric imaging. *Opt. Lett.*, 40(6):882–885, 2015.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[31] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

[32] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016.

[33] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[34] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning Face Representation from Scratch. *ArXiv e-prints*, 2014.

[35] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li. Face matching between near infrared and visible light images. In *ICB*, 2007.

[36] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu. Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces. In *IJCB*, 2017.

[37] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.