TVA-GAN: Attention Guided Generative Adversarial Network For Thermal To Visible Face Transformations

Nand Kumar Yadav, *Student Member, IEEE,* Satish Kumar Singh, *Senior Member, IEEE,* and Shiv Ram Dubey, *Member, IEEE*

***Abstract*—In the recent advancement of machine learning methods for realistic image generation and image translation, Generative Adversarial Networks (GANs) play a vital role. GAN generates novel samples that look indistinguishable from the real images. The image translation using a generative adversarial network refers to unsupervised learning. In this paper, we translate the thermal images into visible images. Thermal to Visible image translation is challenging due to the non-availability of accurate semantic information and smooth textures. The thermal images contain only single-channel, holding only the images' luminance with less feature. We develop a new Cyclic Attention-based Generative Adversarial Network for Thermal to Visible Face transformation (TVA-GAN) by incorporating a new attention-based network. We use attention guidance with a recurrent block through an Inception module to reduce the learning space towards the optimum solution. TVA-GAN is tested and evaluated for thermal to visible face synthesis over the WHU-IIP and Tufts Face Thermal2RGB datasets. The results using the proposed TVA-GAN is promising for face synthesis as compared to the state-of-the-art GAN methods.**

***Index Terms*—GAN, Attention-GAN, Synthesized Loss, Cycle Synthesized Loss, Thermal-Visible Transformation, Thermal-Visible Face Synthesis, Recurrent-Inception module, Attention Block.**

I. Introduction

Visible image generation using thermal images is a very challenging task rather than using Infrared or Near-Infrared images. Near-infrared (NIR) images are close to redlight wavelengths between 700 nm - 1400 nm. NIR images are very close to human vision and discard the color wavelength pieces of information. This results in most articles looking similar to the image converted into gray scale images. Most NIR cameras at night utilizing IR LEDs for illumination are limited in range, usually not more than 500m. While on the other hand, thermal images are far-infrared images with wide-area emission detection. Thermal Infrared (TIR) cameras are sensitive to heat radiation produced by a body. Heat is the electromagnetic waves emitted by a body above the absolute zero temperature, which contains different wavelengths. Both NIR and TIR images capture non-overlapping electromagnetic spectrum. However, thermal images and near-infrared images are very different from each other since thermal images are more specific to capturing images for a particular range of temperature only. Thus, the thermal images(TIR) have more noisy data than the NIR images. So, it is more challenging to generate the actual visible domain images from the corresponding thermal domain images.

In the current scenario of deep learning [1], the image generation tasks handle various applications of computer vision, including image restoration [2], image synthesis [3], face synthesis [4] [5] and many more. We consider the visible face synthesis from the thermal face image as an image-to-image translation problem due to the images' inter-domain transformation. The image-to-image translation [6] method is inspired from the language transformation problem proposed by Mark Twin [7]. Here the language is first transformed from French to English and then back to French, and the final results are compared with the source text string for better translations. The image-to-image translation is effectively handle by Generative Adversarial Networks(GANs) which works on the principle of training a model which learns by balancing false results against true results. With the modern influence of deep learning, different Generative Adversarial Network (GAN) methods [8], [9], [10], [11] have been proposed to deal with the image-to-image translation problems. GAN based models have been also utilized for different applications such as image segmentation [12], image colorization [13], image super-resolution [14], image style transfer [15], and face photo-sketch synthesis [16].

Deep learning methods are prevalent for image-to-image translation in multi-domain scenarios in computer vision, and bio-metrics [1] [17]. The deep learning methods consist of two domains: supervised and unsupervised learning methods. The supervised framework needs tremendous manual work for labeling the data. Generative Adversarial Networks(GANs) have gained massive popularity because of their ability to generate realistic samples within training samples distribution. In proposed TVA-GAN used a thermal face image to feed into the generator network for producing a synthesized real-looking visible face image as the output. The GAN-based image-to-image translation methods comprise two networks: generator and discriminator networks. The discriminator network includes a Convolutional Neural Network (CNN) for two-class classification between the real and fake samples. The generator

N.K. Yadav and S.K. Singh is with the Computer Vision and Biometrics Laboratory at Indian Institute of Information Technology, Allahabad-211015, India (email: nandkmyadav@gmail.com, sk.singh@iiita.ac.in).

S.R. Dubey is with the Computer Vision Group, Indian Institute of Information Technology, Sri City, Chittoor, Andhra Pradesh-517646, India (e-mail: shivram1987@gmail.com, srdubey@iiits.in).

network is an auto-encoder [6][2] that produces high-quality images within the given training set distribution.

The significant commitments of this paper are as follows:

- We propose an Attention-based Generative Adversarial Network (TVA-GAN) for thermal to visible face transformation using an image-to-image translation framework.
- The proposed TVA-GAN's learning space narrowed down towards optimal learning by using attention guidance and the deep feature extraction using the inception network, which helps to learn more local sparse structure and performs better than the traditional methods.
- We proposed a novel generator architecture for TVA-GAN using Recurrent Inception block with attention mechanism to improve the training of Attention network.
- We tested the proposed TVA-GAN for thermal to visible face synthesis using real thermal face images and found improvement over various state-of-the-art methods.

The rest of the paper is described in the following manner: a concise literature review for image translation and thermal to visual transformation is presented in Section II; The proposed TVA-GAN with network analysis and losses are described in Section III; The experimental setup is described in Section IV; The experimental results and observations are described in Section V; and Lastly, the conclusion of the paper is provided in Section VI.

II. Related Work

In the area of methods using machine learning, feature classification using classifiers for recognition task proposed by Jun Li et al. named hallucinating faces using thermal infrared images. In the methods using machine learning, feature classification using classifiers for recognition task proposed by Jun Li et al. [18] named hallucinating faces using thermal infrared images. In comparison, Choi et al. [19] pre-processed the thermal image and normalize the intensity values of images. Choi et al. used Self quotient image(SQI) with the Gaussian filtering (DOG) difference for the recognition task. Cunjian Chen et al. [20] used Pyramid Scale Invariant Feature Transform (PSIFT) for matching the images in thermal and visible domains. These non-deep learning based methods' primary aim is to reduce the domain gap for learning features.

Among deep learning approaches,Vishal M. Patel et al. used polarimetric thermal faces and generative adversarial networks [21] for high-quality visible faces synthesis. The Polarimetric Thermal Database [22] is used in [21] for Face Recognition, which contains polarimetric images with more facial features than actual thermal images. The database consists of only grey channel images, not visible color images. For the same database, Iranmaneshet et al. proposed a Deep Cross Polarimetric Thermal-to-visible face recognition [23] for thermal face recognition. The authors used two CNN and contrastive loss functions to recognize faces from polarimetric and visible domains. Generative Adversarial Networks (GAN) appeared as an unsupervised learning framework for generating the new samples within a given dataset distribution. Different authors proposed different versions of GANs to deal with different problems associated with image generation, translation, and new sample generation. Image-to-image translation methods using GAN proposed by various researchers, which helps translate the images from one domain to another [24],[9],[10],[25],[26],[27],[28]. The ConditionalGAN [29] can be seen as the baseline, which generates new samples with some embedding conditions. The ConditionalGAN generator network can generate samples based on some prior given conditions as class labels. In 2016, first unsupervised image translation network using GAN was proposed by Ming-Yu Liu and Oncel Tuzel, named CoGAN (Coupled Generative Adversarial Network) [30], capable of learning the joint distribution from the marginal distribution of two different domains.

The pix2pix was based on ConditionalGAN and CycleGAN which was quite similar to CoGAN in inter-domain feature learning. The CycleGAN was a state-of-the-art model for the unpaired image to image translations. Its generator was capable of generating more realistic samples than any other methods dealing with unpaired data. The pix2pix used the markovian PatchGAN [8] discriminator network, and it displayed promising results for the paired image transformation. The pix2pix restricted for paired image transformation using the same set of images in different domains. pix2pix used the PatchGAN discriminator for labeling the generated image patches. The paired image dataset collection is expensive and suffers from long procedural processes. To remove such problems associated with pix2pix, CycleGAN proposed, which can transform the inter-domain images without having paired datasets. CycleGAN converts the source domain images into the target domain images of the same semantic information. Further network converts them back to the source domain images. Which helps to decreasing the divergence of the learning space and increasing the quality of generated images. On the other side, Yi et al. proposed DualGAN [10] similar to CycleGAN for the image-to-image translation, which varies from CycleGAN in terms of the loss functions. The DualGAN exercises reconstruction loss, whereas the CycleGAN practices the Cycle-consistency loss. In most of the incidents, the CycleGAN outperforms the DualGAN. Thus, we use the CycleGAN framework in the proposed model.

In a recent development, Self-Attention GAN is proposed [28], which is also known as an intra-attention network capable of boosting the CNN performance because the attention network focuses more on the essential features of the images. Self-Attention GAN learns the long-range multi-level dependencies by attending the response at a specific position of images. The attention-based networks help to eliminate intense training of deep neural networks compared to CNN models [28] [31] [32]. Recently, the attention-based networks are also proposed by Mejjati et al. [26] and Tang et al. [33] for image-to-image translation using GANs. Both of these methods used attention guided generator for the foreground image generation and preserved the background information using inverse mapping of generator output and concatenated them in final synthesizing. There is few more attention-based GANs for image-to-image translation, including Multi-channel Attention GAN [34] and Deep-Attention GAN [35]. Attribute

guided GAN [36] is proposed for sketch generation. Attention-based two-stream CNNs [37], [32] are proposed for spoofing detection in faces.

III. Proposed TVA-GAN Model

n this section, we present the proposed Thermal to Visible transformation Attention Guided Generative Adversarial Network (TVA-GAN) for Thermal to Visible face synthesis. The proposed TVA-GAN architecture is illustrated in Fig. 3. We use the paired dataset $A_{j=1}^{n} = (X_j, Y_j)_{j=1}^{n}$, $x \in X$ and $y \in Y$, where x_j and y_j are the pairs of thermal and corresponding visible images. We use CycleGAN [9] framework with U-Net [38] based architecture. The generator network consists of an encoder and a decoder. The encoder is based on the Recurrent-Inception modules and the decoder is based on the attention mechanisms. The proposed TVA-GAN translates the images from source domain (x) to target domain (y) and target domain (y) to source domain (x) in cyclic manner. We use two Attention Guided Generator Networks, i.e., G_{xy} to translate images from domain x to domain y ($x \rightarrow y$) and G_{yx} to generate the image in domain x from domain y ($y \rightarrow x$). The generator network used in the proposed TVA-GAN has an inbuilt attention mechanism .

The proposed TVA-GAN method trained end to end using the various types of loss functions. For better convergence, we combined multiple losses to add different curvatures in the optimization. The followings are the losses used in this paper: Adversarial loss, Cycle loss, Synthesized loss, Cycle synthesized loss, Feature reconstruction loss (i.e., perceptual loss)

Attention Block: We use attention gates [39] as Attention block in our proposed network to capture sizeable receptive field and semantic contextual information. While applying multi-stage CNN, the attention gate reduces the feature responses for irrelevant background regions. There is no restriction for cropping an ROI (region of interest) between the network layers. Attention gate output is obtained from element-wise multiplication between input feature maps denoted as $z_k$and q_k^{att} respectively.

z^k is the feature map of k$^{\text{th}}$ layer in CNN network. $z_j^k \in \mathbb{R}^{F_k}$ where F_k represents the number of feature maps in k^{th} layer. Attention gate helps to focus on subset of a specific region of target structure. The gating vector denoted by g_j, helps to analysing spatial regions by providing contextual and activation information. Where $g_j \in \mathbb{R}^{F_g}$ used for determining the focus region of pixel j. In the attention block ReLU presented by σ_1.

$$\sigma_1(z_j^k) = max(0, z_j^k, c)$$

We use additive attention, where the attention map calculated between previous up-sampling layer and corresponding down-sampling layer of encoder block in network. Hence both layers attention map added and perform operation for getting q_{att}^k. Both the vectors after channel wise convolution of 1 summed element wise because it shows better results than multiplicative attention [40](element wise multiplication increases the network complexity).

$$q_{att}^k = \sigma_2(\varphi^T(\sigma_1(W_z^T z_j^k + W_g^T g_j + b_g)) + b_\varphi)$$
$$\hat{z}_j^k = (q_{att}^k * z^k))$$

where $\sigma_2(z_{j,c}) = \frac{1}{1+exp(-z_{j,c})}$ represents the Sigmoid activation function where j and c denotes the spatial and channel dimensions. $W_z \in \mathbb{R}^{F_k \text{x} Fint}$, $W_g \in \mathbb{R}^{F_g \text{x} Fint}$ and $\varphi \in \mathbb{R}^{F_{int} \text{x1}}$ represent the linear transformation. F_{int} denotes the no of output channel for each 1×1 convolution, and $b_\varphi \in \mathbb{R}$ and $b_g \in \mathbb{R}^{F_{int}}$ represent the bias term. In brief, two input feature maps passed through the $1 \times 1 \times 1$ channel-wise convolution after that combined through adding the outputs and pass by ReLU activation. Therefore second channel-wise convolution was performed using $1 \times 1 \times 1$ kernels and passed through the Sigmoid layer to obtain the mask and concatenate the attention mask with up-sampled feature maps. Attention Block shown in Fig. 1.

Note: The linear transformations are computed by $1 \times 1 \times 1$ channel-wise convolutions. Attention block described in Table II

Recurrent Inception Block For better learning of the contextual information, we used recurrent block with $t = 2$ occurrences. In the proposed RCIN, the recurrent block results in more network depth with fewer parameters and learning by weight sharing. For learning the globally as well locally, we used the inception network with the recurrent network. Inception also helps to make networks computationally cheaper in terms of parameters. While using two recurrent blocks together, we found a large no of computational parameters besides this. We used a novel recurrent inception module that reduces parameters and learns both locally and globally due to large and small filter sizes (3×3, 5×5 and 1×1) . We pass each layer through ReLU layer(except the max-pooling layer), as shown in Fig.1. To overcome the problem of vanishing gradients. The ReLU used in architecture advantages with faster and more efficient learning due to no error while back-propagating the gradients in the network with fewer computational parameters than softmax. To make the network smaller, we fixed the no of output filters for 5×5 most immense kernel size in inception block kernel size; the no of output filters fixed to 16 instead of deriving from input parameters, because filters derived from input parameters results in more number filter layers introduced in the network and increases the network complexity. The recurrent inception block architecture described in Table I.

Adversarial Loss: Adversarial loss measures the error for generator and discriminator networks. The generator network generates the fake image specimens. The discriminator network produces labels for the generated image samples as fake/real, depending upon how each generated image data distribution matches the corresponding real image data distribution. The vanilla GAN uses negative log-likelihood loss [41], which leads to instability in training. To overcome the instability problem the proposed TVA-GAN model uses LSGAN [42].

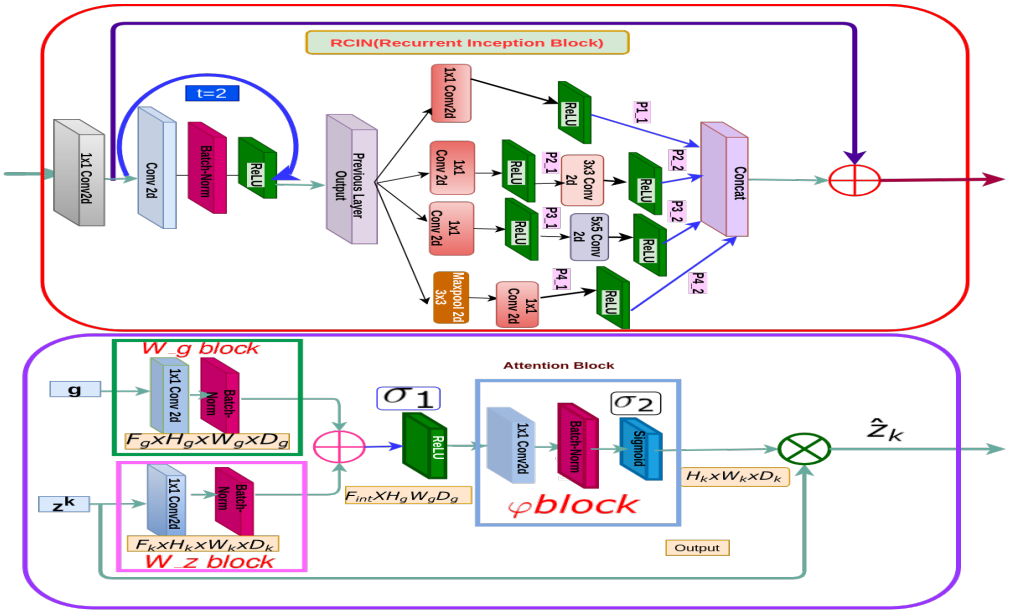

Fig. 1: A network block of Recurrent Inception Block (RCIN). Attention mechanism for network architecture is attached in attention block.

The GAN adversarial loss for $X \rightarrow Y$ transformation is described as below where G_{xy} denotes the generator function for transforming the images domain x to domain y. While D_Y is discriminator function for domain Y.

$$\mathcal{L}_{GAN}(G_{xy}, D_Y) = Min_{G_{xy}} Max_{D_Y} = \mathbb{E}_{y \sim p_{data(y)}}[(D_Y(y) - 1))^2] + \mathbb{E}_{x \sim p_{data(x)}}[(D_Y(G_{xy}(x)) - 1)^2]$$

where $x \in X$ and $y \in Y$. Similarly, GAN adversarial loss computed for $Y \rightarrow X$ transformation ($\mathcal{L}_{GAN}(G_{yx}, D_X)$) . Where G_{yx} denotes the generator function for transforming the images domain y to domain x. While D_X is discriminator function for domain X.

$$\mathcal{L}_{GAN}(G_{yx}, D_X) = Min_{G_{yx}} Max_{D_X} = \mathbb{E}_{x \sim p_{data(x)}}[(D_X(x) - 1))^2] + \mathbb{E}_{y \sim p_{data(y)}}[(D_X(G_{yx}(y)) - 1)^2]$$

Cycle Loss: We use cycle-consistency loss (cycle loss) [9] in the objective function of the proposed method. It is computed using the L_1 distance between the real image and the cyclic reconstructed image in both forward and backward transformations. The forward cycle loss is defined as,

$$\mathcal{L}_{Cyc_F} = ||x - G_{yx}(G_{xy}(x))||_1$$

Similarly, the backward cycle loss is computed as,

$$\mathcal{L}_{Cyc_B} = ||y - G_{xy}(G_{yx}(y))||_1$$

where $x \in X$ and $y \in Y$.

Cycle-Synthesized Loss: The cycle-synthesized loss [43] is used in the proposed model to make training better. We calculate the cycle-synthesized loss as L_1 loss between the cycled/reconstructed image and the synthesized image in cross-domains. The cycle-synthesized losses are computed as,

$$\mathcal{L}_{Csl_1} = ||G_{xy}(G_{yx}(y) - G_{xy}(x)||_1$$
$$\mathcal{L}_{Csl_2} = ||G_{yx}(G_{xy}(x) - G_{yx}(y)||_1$$

where $G_{yx}(y)$ and $G_{xy}(x)$ are the synthesized images and $G_{yx}(G_{xy}(x)$ and $G_{xy}(G_{yx}(y)$ are the cycled images.

Synthesized Loss: Synthesized loss is calculated between the generated image and the input image without using the detachment of computation graph, which helps to back-propagate the network loss. For $A \in X$ and $B \in Y$ the synthesized losses in domains A and B are defined as,

$$\mathcal{L}_{Sl_A} = ||x - G_{yx}(y)||_1$$
$$\mathcal{L}_{Sl_B} = ||y - G_{xy}(x)||_1.$$

Feature Reconstruction Loss: We estimate the loss for related feature representation between the target image and the generated image. The same is also performed for the target and the corresponding reconstructed image. We use mean square error to compute the distance between the extracted features, where feature extraction performed using the pre-trained VGG-19 network used in the perceptual loss. For any trained network ψ, let $\psi_k(y)$ represents the activation feature map of dimension $W_k \times H_k \times C_k$ corresponding to the k_{th} convolution layer. Where C represents a number of channel, W width of input image and H height of input image, ψ is the pre-trained VGG-19 model. While processing image y through pre-trained network's (ψ) k$^{\text{th}}$ layer we get the feature map $\psi_k(y)$.

$$l^{\psi,k}_{feat}(\hat{y}, y) = \frac{1}{W_k H_k C_k} ||\psi_k(\hat{y}) - \psi_k(y)||_2^2$$

where y and $\hat{y}$ are the original and the generated images, respectively. Using the above function, we compute the fol-

TABLE I: **Recurrent Inception Block**

Recurrent Inception Block (input_ch = in_ch, output_channels)				
Layers	**kernel_size**	**stride**	**Padding**	**channels in,out**
C1 = Conv2d	1	1	0	in_ch,in_ch
Conv2d + BatchNorm + ReLU	3	1	1	in_ch,in_ch
Conv2d_Inception(1_1)	1	1	-	in_ch,in_ch/4
ReLU				
Conv2d_Inception(2_1)	1	1	-	in_ch,in_ch
ReLU				
Conv2d_Inception(2_2)	3	1	1	in_ch,in_ch/4
ReLU				
Conv2d_Inception(3_1)	1	1	-	in_ch,16
ReLU				
Conv2d_Inception(3_2)	5	1	2	16,in_ch/4
ReLU				
Maxpool2d_Inception(4_1)	3	1	1	–
Conv2d_Inception(4_2)	1	1	-	in_ch,in_ch/4
ReLU				
C = Conat(1_1),(2_2),(3_2),(4_2)				
(**C1 + C**) ,output_channels = input_channels				

TABLE II: **Attention Block**

W_g block, input= in			
Layers	**kernel_size**	**stride**	**channels in,out**
Conv2d + BatchNorm	1	1	in,in/2
W_z block, input= in			
Layers	**kernel_size**	**stride**	**channels in,out**
Conv2d + BatchNorm	1	1	in,in/2
A = ReLU(output(W_z) + output(W_g))			
φ block, input= in/2			
Layers	**kernel_size**	**stride**	**channels in,out**
Conv2d + BatchNorm	1	1	in,in/in
Sigmoid			
Out = φ(A) * input((W_z))			

lowing feature reconstruction losses where $x \in X$ and $y \in Y$:

$$\mathcal{L}_{real}^{fake}(A) = l_{feat}^{\psi,k}(x, G_{yx}(x))$$
$$\mathcal{L}_{real}^{fake}(B) = l_{feat}^{\psi,k}(y, G_{xy}(y))$$
$$\mathcal{L}_{real}^{recon}(A) = l_{feat}^{\psi,k}(x, G_{xy}(G_{yx}(x)))$$
$$\mathcal{L}_{real}^{recon}(B) = l_{feat}^{\psi,k}(y, G_{yx}(G_{xy}(y)))$$
$$\mathcal{L}_{fake}^{recon}(A) = l_{feat}^{\psi,k}(G_{xy}(y), G_{xy}(G_{yx}(x)))$$
$$\mathcal{L}_{fake}^{recon}(B) = l_{feat}^{\psi,k}(G_{yx}(x), G_{yx}(G_{xy}(y)))$$

Objective Function: The final objective function for the proposed TVA-GAN is given as follows:

$$\mathcal{L}(G_{xy}, G_{yx}, D_X, D_Y) = \mathcal{L}_{GAN} + \mathcal{L}_{Cyc} + \mathcal{L}_{Csl} + \mathcal{L}_{Sl} + \mathcal{L}_{FR}$$

where

$$\mathcal{L}_{GAN} = (\mathcal{L}_{GAN}(G_{xy}, D_Y) + \mathcal{L}_{GAN}(G_{yx}, D_X))$$
$$\mathcal{L}_{Cyc} = \lambda_{Cyc}(\mathcal{L}_{Cyc_F} + \mathcal{L}_{Cyc_B})$$
$$\mathcal{L}_{Csl} = \lambda_{Csl}(\mathcal{L}_{Csl_1} + \mathcal{L}_{Csl_2})$$
$$\mathcal{L}_{Sl} = \lambda_{Sl}(\mathcal{L}_{Sl_A} + \mathcal{L}_{Sl_B})$$
$$\mathcal{L}_{FR} = \lambda_{feat}(\mathcal{L}_{real}^{fake}(A) + \mathcal{L}_{real}^{fake}(B) + \mathcal{L}_{real}^{recon}(A) + \mathcal{L}_{real}^{recon}(B) + \mathcal{L}_{fake}^{recon}(A) + \mathcal{L}_{fake}^{recon}(B))$$

where λ is the weight hyperparameters for different type of losses.

TABLE III: **Generator Network Architecture**

Layers	**kernel_size**	**stride**	**Padding**	**channels in,out**
Encoding Block				
R1 = Recurrent Inception Block (in_channels =3, out_channels =64)				
AvgPool2d	2	2	-	-
R2 = Recurrent Inception Block (in_channels =64, out_channels =128)				
AvgPool2d	2	2	-	-
R3 = Recurrent Inception Block (in_channels =128, out_channels =256)				
AvgPool2d	2	2	-	-
R4 = Recurrent Inception Block (in_channels =256, out_channels =512)				
AvgPool2d	2	2	-	-
R5 = Recurrent Inception Block (in_channels =512, out_channels =1024)				
Decoding +Concatenation				
U5 = Upsample(scale_factor = 2.0) + Conv2d + BatchNorm + ReLU	3	1	1	1024,512
A4 =Attention_block (U5,R4)				
C5 = Concat(A4,U5)				
Recurrent Inception Block(C5)				
U4 = Upsample(scale_factor = 2.0) + Conv2d + BatchNorm + ReLU	3	1	1	512,256
A3 =Attention_block (U4,R3)				
C4 = Concat(A3,U4)				
Recurrent Inception Block (C4)				
U3 = Upsample(scale_factor = 2.0) + Conv2d + BatchNorm + ReLU	3	1	1	256,128
A2 =Attention_block (U3,R2)				
C3 = Concat(A2,U3)				
Recurrent Inception Block(C3)				
U2 = Upsample(scale_factor = 2.0) + Conv2d + BatchNorm + ReLU	3	1	1	128,64
A1 =Attention_block (U2,R1)				
C2 = Concat(A1,U2)				
Recurrent Inception Block(C2)				
Conv2d	1	1	0	64 ,3
tanh				

TABLE IV: **Discriminator Network Architecture**

Layers	**Padding**	**Stride**	**Output**
Conv2d + LeakyReLU	1	2	(64, 128, 128)
Conv2d + LeakyReLU + Instance Norm	1	2	(128, 64, 64)
Conv2d + LeakyReLU + Instance Norm	1	2	(256, 32, 32)
Conv2d + LeakyReLU + Instance Norm	1	2	(512, 16, 16)
Conv2d + LeakyReLU + Instance Norm	1	2	(512, 8, 8)
Conv2d + LeakyReLU + Instance Norm	1	1	(512, 7, 7)
Conv2d	1	1	(1, 6, 6)

IV. Experimental Setup

A. Network Architecture

For training the network we use newly proposed recurrent inception block with attention networks. The integration of recurrent inception block with attention networks makes it better for learning in image-to-image translation task. We use CycleGAN network as the base model for translation task. The proposed method can generate more realistic and accurate translation task while synthesising the images. The proposed method contains two Generator networks (i.e., G_{xy}, and G_{yx}) and two Discriminator networks (i.e., D_Y and D_X) for both domains, respectively. The generator has inbuilt

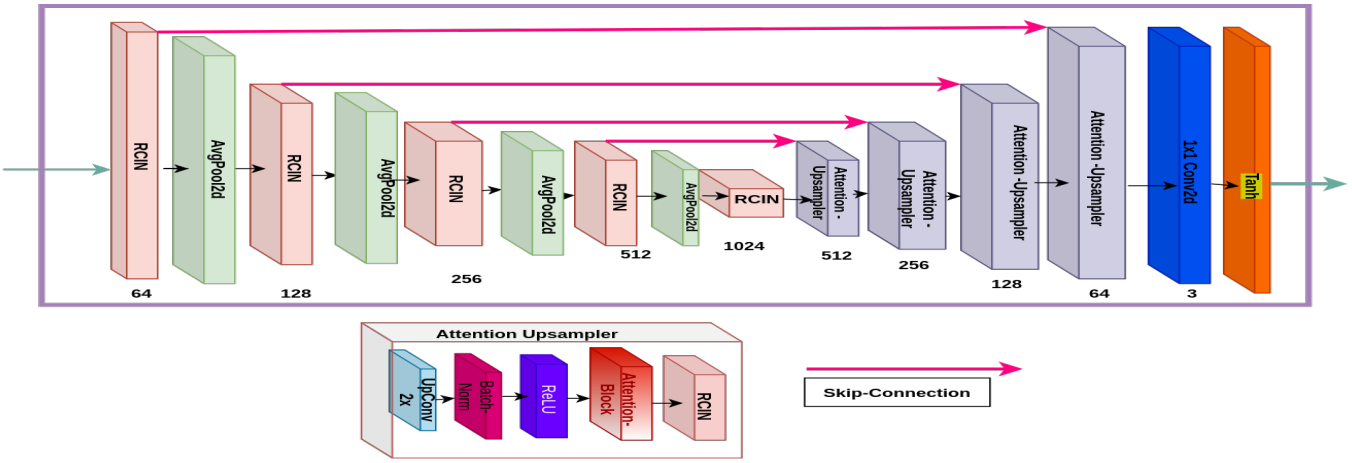

Fig. 2: Proposed TVA-GAN network architecture where RCIN denotes recurrent inception block.

attention mechanisms. Attention network was proposed by Goodfellow et al. for handling the long-range dependencies in the network [28]. It also helps to the proposed TVA-GAN to handle the background information without introducing any new network. We follow the architecture having an integrated attention module to take care of long-range dependencies.

Generator Network: We use the recurrent-inception attention-based network architecture in this paper in the generator network. The encoder of generator network includes recurrent-inception block as examined in Table I. The recurrent-inception helps to improve the network performance and the learning of optimal local sparse structure. The attention block consists of the Attention-Gate [39] architecture outlined in Table II. The attention block is used in the decoder only after every up-sampling layer, followed by the Convolutional layer combined with batch normalization and ReLU activation function. The attention-block finds the scalar attention value for each pixel vector by additive attention learned through linear transformation using $1\times1\times1$ channel wise convolutions. The generator architecture summary is presented in Table III.

Discriminator Network: In the discriminator network architecture, we use the PatchGAN discriminator proposed in pix2pix, known as Markovian Patch-GAN discriminator with five-layer architecture. We feed the discriminator network with 256×256 images generated by the generator network. Discriminator's 1^{st} layer is a convolution layer with LeakyReLU activation function. After that, each convolution layer is followed by the instance-normalization and LeakyReLU activation function. We use 4×4 kernel in each Convolutional layer with stride 2 and padding 1. Last layer of architecture contains only convolution layer. The network architecture of discriminator network is summarized in Table IV.

B. Baseline Methods

The proposed TVA-GAN for Thermal-Visible synthesis is compared with current baseline methods of image-to-image translation by following its original settings.

1) pix2pix [24]: pix2pix is used for paired image dataset translates the images from one domain to another using the U-net generator network with the PatchGAN discriminator network. It works based on conditional data input. Original settings used for evaluation of network performance. [1]

2) CycleGAN [9]: CycleGAN is proposed for the unpaired image-to-image translation method by using cycle-consistency loss. It transforms the source domain image into the target domain image and then reconstructs the target domain image to the source domain image. The cycle-consistency loss is calculated between the source image and reconstructed image.

3) DualGAN [10]: DualGAN also refers to nearly the same methodology as CycleGAN, but uses reconstruction loss rather than cycle-consistency loss. Also, it does not require the paired data in the image translation task. DualGAN, with its original setting, is used for performance evaluation. [2]

4) PCSGAN [25]: PCSGAN also refers to nearly the same methodology as CycleGAN, but uses cycle perceptual loss with synthesized perceptual loss rather than cycle-consistency loss. It uses the paired data in the image translation task.

5) AGGAN [26]: An attention-guided model (AGGAN), proposed by Mejjati et al., extracts the attention map to find the foreground and background of images. The attention mechanism discovers the region of translation in the opposite domain by finding the attention map. [3]

6) AttentionGAN [27]: AttentionGAN practices the same mechanism introduced in CycleGAN with an inbuilt attention mechanism to find an attention mask with content mask to transform the images from one domain to another. [4]

C. Datasets Used

We test our model for two thermal-visible datasets, namely WHU-IIP and Tufts Face Thermal2RGB; both datasets contain the thermal and real visible face pairs. We use the WHU-IIP [44] and Tufts Face Thermal2RGB [45] datasets for thermal to visible face synthesis using the proposed TVA-GAN method

[1] https://github.com/junyanz/pytorch-CycleGAN-and-Pix2pix
[2] https://github.com/duxingren14/DualGAN
[3] https://github.com/AlamiMejjati/Unsupervised-Attention-guided-Image-to-Image-Translation
[4] https://github.com/Ha0Tang/AttentionGAN

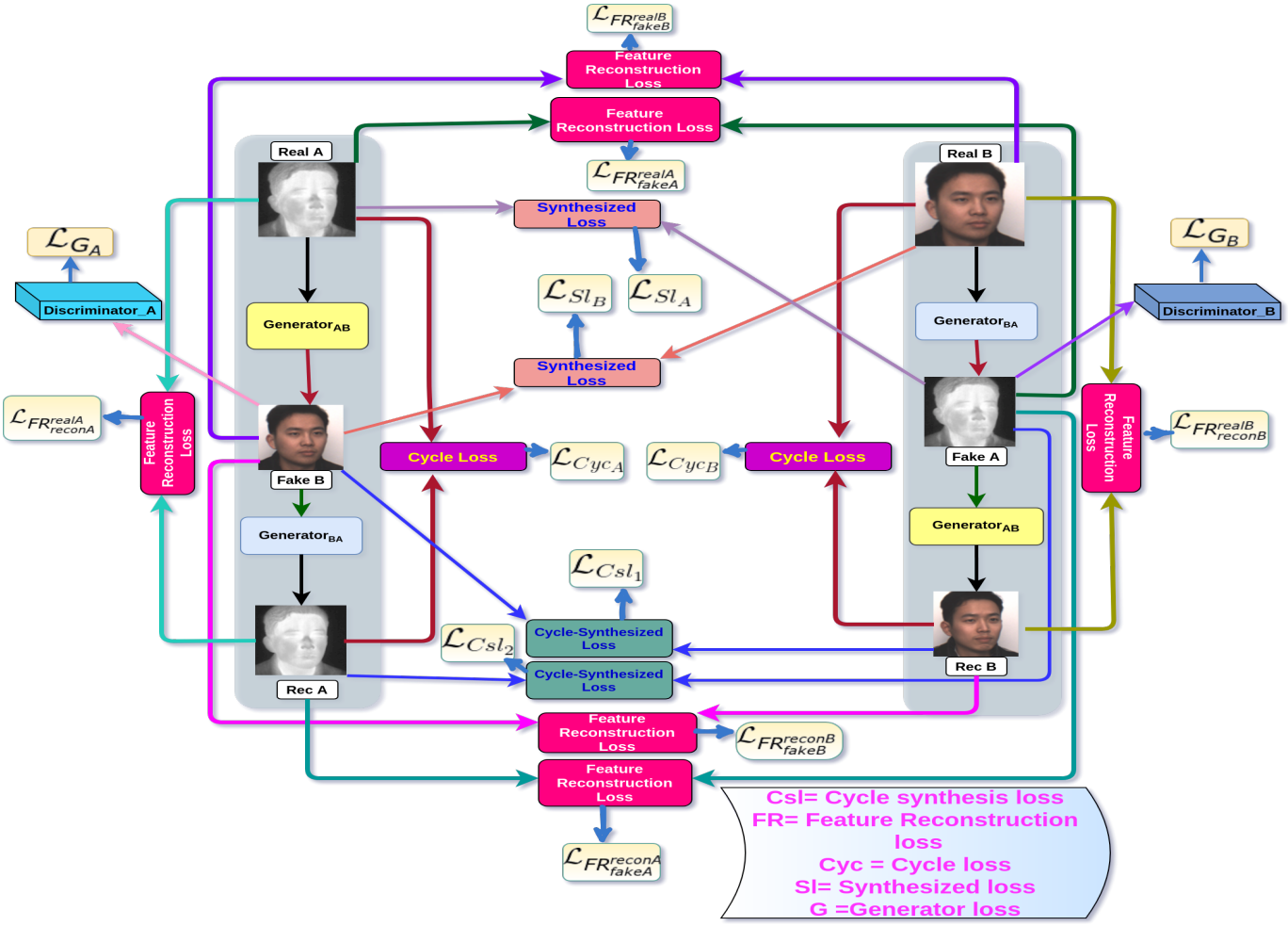

Fig. 3: Proposed TVA-GAN Architecture.

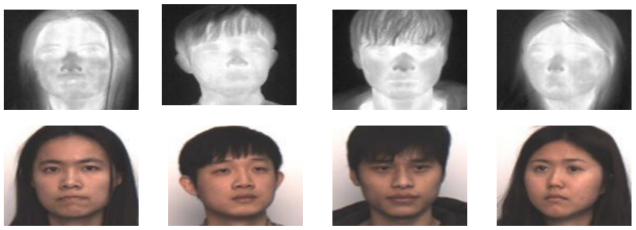

Fig. 4: WHU-IIP dataset samples of face images in thermal domain (top) and visible domain (bottom)

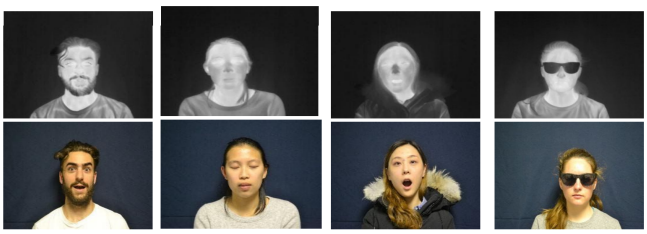

Fig. 5: Tufts Face Thermal2RGB dataset samples of face images in thermal domain (top) and visible domain (bottom)

and existing GAN based methods. For WHU-IIP for thermal to real visual transformation, 552 training image pairs, and 240 testing image pairs are considered in the experiments. We use 403 images for training and 156 images for testing in paired manner for Tufts Face Thermal2RGB dataset. Tufts Face thermal2RGB dataset contains more diverse data than WHU-IIP to judge the generalization capability of the proposed model. It includes images of people having various races with different facial attributes, including some people who have sunglasses and spectacles.

D. Parameter Settings

For all the datasets used for training and testing, the images are resized to the dimensions as $256 \times 256 \times 3$ (where 3 denotes the no. of channels). Similar to CycleGAN, pool size is set to 50. We use diffGrad optimizer [46] for the proposed TVA-GAN because previously proposed optimizers [47], [48] suffer from adjustment of learning-rate update. For the pix2pix method, we use the batch normalization based on the original implementation. For the CycleGAN and DualGAN, we use the batch normalization method as proposed in the original network for comparison with our results. We use lsgan loss [42] as used in CycleGAN for training stability of the proposed model through out the training process. The loss weight hyperparameters used in the final objective function are listed in Table VII. We use the diffGrad optimizer with a learning rate of 0.0002 and momentum terms $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The linear decay is used to reduce the optimizer's learning rate

TABLE V: The quantitative results comparison over WHU-IIP dataset of the proposed TVA-GAN model with recent state-of-the-art GAN models. Note that the higher value is better for SSIM and PSNR, whereas lower value is better for LPIPS and VGG-FaceLoss.

Method	SSIM	LPIPS	PSNR	VGG-FaceLoss
pix2Pix	0.7709	0.087	29.52	0.6176
CycleGAN	0.7573	0.084	29.50	0.6271
DualGAN	0.7623	0.080	29.42	0.5887
PCSGAN	0.8163	0.063	30.08	0.5160
AGGAN	0.7423	0.085	29.02	0.6411
AttentionGAN	0.6368	0.115	28.79	0.8021
TVA-GAN(ours)	**0.8444**	**0.052**	**29.96**	**0.4756**

TABLE VI: The quantitative results comparison over Tufts FaceThermal2RGB dataset of the proposed TVA-GAN model with recent state-of-the-art GAN models. Note that the higher value is better for SSIM and PSNR, whereas lower value is better for LPIPS and VGG-FaceLoss.

Method	SSIM	LPIPS	PSNR	VGG-FaceLoss
pix2Pix	0.5027	0.231	28.36	0.5980
CycleGAN	0.5805	0.182	28.62	0.7832
DualGAN	0.5652	0.219	28.77	0.7684
PCSGAN	0.6244	0.127	31.02	0.5569
AGGAN	0.5876	0.188	28.76	0.8227
AttentionGAN	0.5534	0.212	28.54	0.8092
TVA-GAN(ours)	**0.6924**	**0.048**	**31.52**	**0.3321**

till 0. We update the learning rate after every 50 epochs. The non-attention-based methods are trained for 200 epochs. The attention-based methods, like AGGAN and AttentionGAN, are trained for 100 and 60 epochs, respectively, as per the source paper code. The proposed TVA-GAN model is trained for 200 epochs. The proposed method converges in fewer epochs (i.e., 100) for WHU-IIP dataset while requires 200 epochs for complex Tufts Face Thermal2RGB dataset. We train the proposed method for 200 epochs for both datasets.

E. Evaluation Metrics

For the quantitative analysis of our results as compared to the state-of-the-art methods, we use SSIM [49], LPIPS[50] , PSNR [49] and VGG-FaceLoss evaluation metrics. The Structural Similarity Index (SSIM) is used to measure the structural similarity between the generated and real visible face images. SSIM shows better human-level visual perception. Higher SSIM means close structural similarity between the generated image and the actual visible face image. Peak Signal-to-Noise Ratio (PSNR) is computed to measure the quality of generated images. Learned Perceptual Image Patch Similarity (LPIPS) helps to find the patch level similarity as we use the PatchGAN discriminator. This evaluation helps to understand the quality of generated images using the proposed method. We also compute VGG-FaceLoss to ensure feature-level similarity. It uses a pre-trained VGGFace to extract the features from a synthesized face image and actual visible face image and computes the L1 distance between them. We also use Visual Information Fidelity (VIF) [51] to study the proposed method using different losses. VIF is used to compare the visual information among the reference image and generated image. The VIF helps to distinguish the generated images from the reference images as human visual system does. So, VIF helps to understand how accurate transformation occurs while our proposed method transforms the thermal images into visible images.

TABLE VII: Training parameter values used for different losses.

Notation	Value
λ_{Cyc}	10
λ_{feat}	1
λ_{Sl}	15
λ_{Csl}	1 for Tufts, 0 for WHU-IIP

TABLE VIII: Losses notations used in the proposed TVA-GAN model.

Loss	Notification
Adversarial Loss	AL
Cycle Loss	Cyc
Synthesized Loss	Sl
Cycle-Synthesized Loss	Csl
Feature Reconstruction Loss	FR

TABLE IX: The quantitative results comparison of the proposed TVA-GAN model using various Losses for WHU-IIP dataset. The higher value is better except for LPIPS.

Method	SSIM	VIF	PSNR	LPIPS
AL	0.5245	0.7817	28.37	0.196
AL+Cyc	0.7664	0.8298	29.30	0.083
AL+Cyc+Sl	0.8290	0.8341	29.92	0.058
AL+Cyc+Sl+FR	0.8444	0.8343	29.96	0.052
AL+Cyc+Sl+FR+Csl	**0.8444**	**0.8343**	**29.96**	**0.052**

TABLE X: The quantitative results comparison of the proposed TVA-GAN model using various Losses for Tufts Face Thermal2RGB dataset. The higher value is better except for LPIPS.

Method	SSIM	VIF	PSNR	LPIPS
AL	0.4963	0.7829	29.10	0.198
AL+Cyc	0.4867	0.7769	28.87	0.219
AL+Cyc+Sl	0.6813	0.8056	31.46	0.070
AL+Cyc+Sl+FR	0.6905	0.8044	31.35	0.050
AL+Cyc+Sl+FR+Csl	**0.6924**	**0.8083**	**31.52**	**0.048**

V. Experimental Results and Observations

A. Quantitative Result Analysis

The proposed TVA-GAN generates more realistic and natural-looking images while transforming the thermal domain into the visual domain. TVA-GAN shows more promising results than the state-of-the-art attention and non-attention-based GAN models.

The proposed method compared with recent state-of-the-art attention-based method AGGAN[26] and AttentionGAN[27], as well as non-attention-based method as pix2pix[24],CycleGAN[9], DualGAN[10], PCSGAN[25].

For thermal to visual synthesis, the quantitative results of TVA-GAN concerning various state-of-the-art methods are reported in Table V for the WHU-IIP dataset and Table VI for the Tufts Face Thermal2RGB dataset. We found that TVA-GAN performs better than all state-of-art methods in terms of the SSIM, LPIPS, and VGG-FaceLoss for both WHU-IIP and Tufts Face Thermal2RGB datasets. It's performance is slightly low in terms of PSNR compared to PCSGAN for the WHU-IIP dataset.

- The gain in term of % for SSIM score using WHU-IIP dataset, as reported in Table V, is {9.55%, 11.50%, 10.77%, 3.44% 13.75%, 32.60%} higher than non-attention-based methods such as pix2pix, CycleGAN, DualGAN, PCSGAN and attention-based methods such as AGGAN and AttentionGAN, respectively.

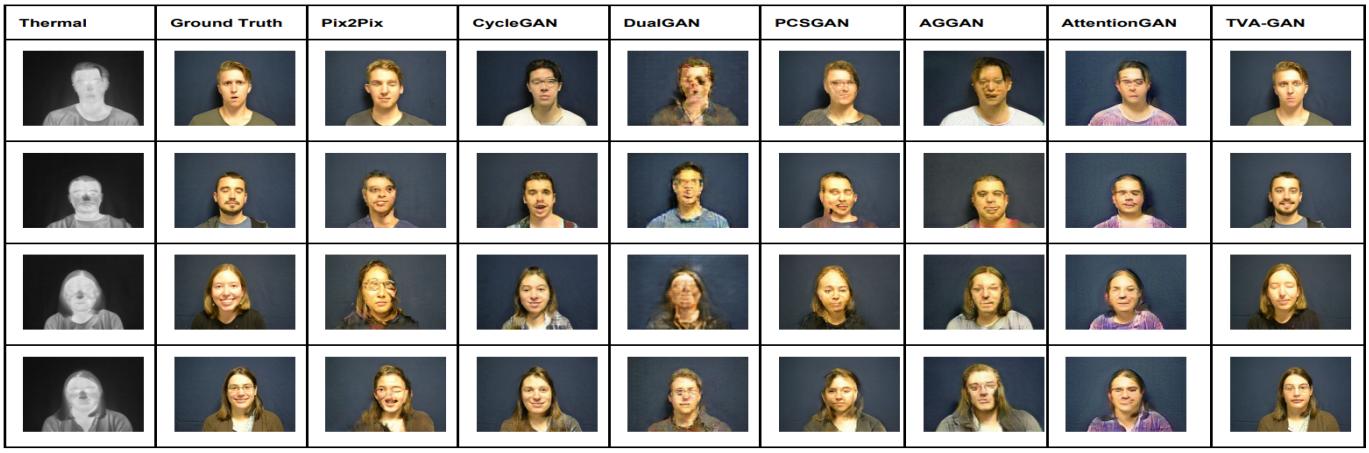

Fig. 6: Qualitative comparison for Thermal to Visible domain transformation using Tufts Face Thermal2RGB dataset. From left to right: Thermal images, corresponding Ground Truth images, images generated using pix2pix, CycleGAN, DualGAN, PCSGAN, AGGAN, AttentionGAN and TVA-GAN models. The TVA-GAN generates more realistic and fair images.

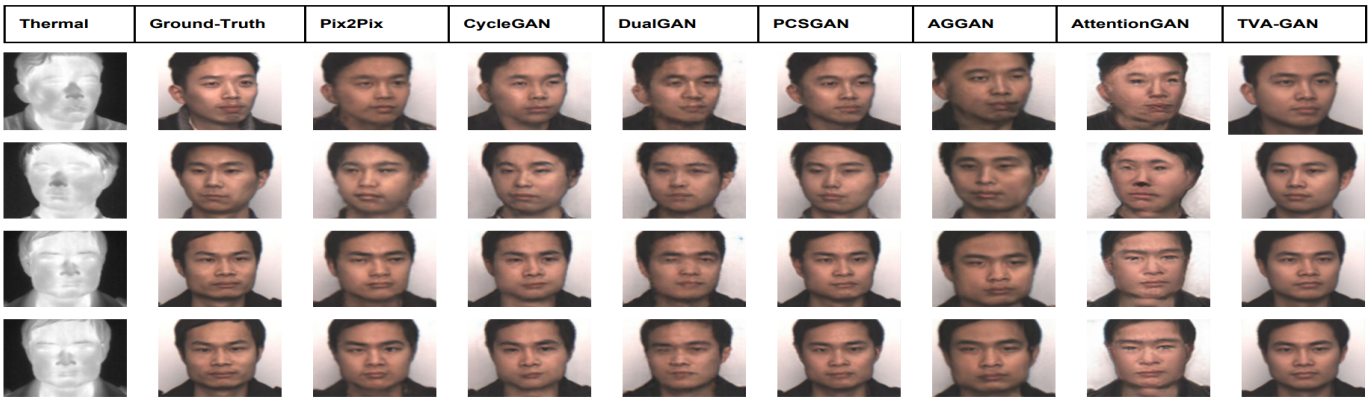

Fig. 7: Qualitative comparison for transformation of Thermal to Visible using WHU-IIP face dataset. From left to right: Thermal images, corresponding Ground Truth images, images generated using pix2pix, CycleGAN, DualGAN, PCSGAN, AGGAN, AttentionGAN and TVA-GAN models. The TVA-GAN generates more realistic and fair images.

- The gain in term of % for PSNR score using WHU-IIP dataset, as reported in Table V, is $\{1.49\%, 1.56\%, 1.84\%, -0.39\%, 3.24\%, 4.06\%\}$ higher than non-attention-based methods such as pix2pix, CycleGAN, DualGAN, PCSGAN and attention-based methods such as AGGAN and AttentionGAN, respectively.
- The gain in term of % for SSIM score using Tufts dataset, as reported in Table VI, is $\{37.74\%, 19.28\%, 22.51\%, 10.89\%\ 17.84\%, 25.12\%\}$ higher than non-attention-based methods such as pix2pix, CycleGAN, DualGAN, PCSGAN and attention-based methods such as AGGAN and AttentionGAN, respectively.
- The gain in term of % for PSNR score using Tufts Face Thermal2RGB dataset, as reported in Table VI, is $\{11.14\%, 10.13\%, 9.56\%\ , 1.61\%\ 9.60\%\ , 10.44\%\}$ higher than non-attention-based methods such as pix2pix, CycleGAN, DualGAN, PCSGAN and attention-based methods such as AGGAN and AttentionGAN, respectively.

On the other hand, the proposed TVA-GAN shows lower score for LPIPS and VGG-FaceLoss for both WHU-IIP and Tufts Face Thermal2RGB datasets.

- The proposed TVA-GAN shows reduction for LPIPS in terms of %, as reported in Table V, for WHU-IIP dataset by $\{40.23\%, 38.01\%, 35.00\%, 17.46\%\ 38.82\%, 54.78\%\}$ than pix2pix, CycleGAN, DualGAN, PCSGAN, AGGAN, and AttentionGAN, respectively.
- The proposed TVA-GAN shows reduction for LPIPS in terms of %, as reported in Table VI, for Tufts Face Ther-

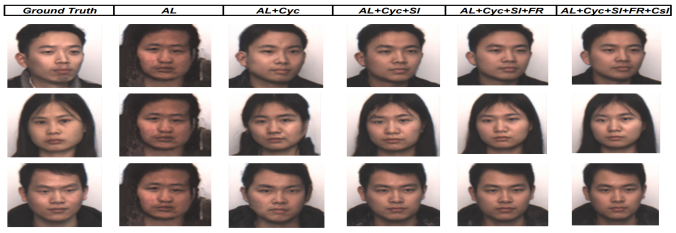

Fig. 8: Qualitative comparison for thermal to visible transformation using WHU-IIP dataset. From left to right: Ground Truth, AL, AL+Cyc, AL+Cyc+Sl, AL+Cyc+Sl+FR, AL+Cyc+Sl+FR+Csl. The TVA-GAN using all the losses generates more realistic and fair images.

Fig. 9: Qualitative comparison for thermal to visible transformation using Tufts Face Thermal2RGB dataset. From left to right: Ground Truth, AL, AL+Cyc, AL+Cyc+Sl, AL+Cyc+Sl+FR, AL+Cyc+Sl+FR+Csl. The TVA-GAN using all the losses generates more realistic and fair images.

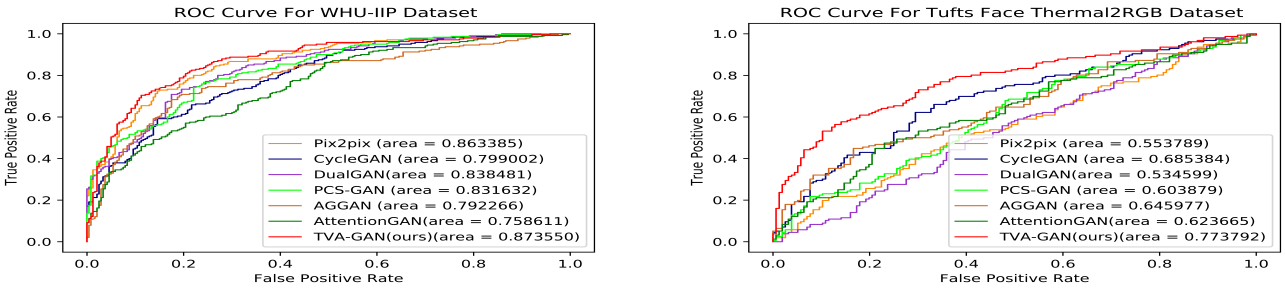

Fig. 10: ROC Curves for Face Verification using DeepFace with cosine-similarity as distance metric over WHU-IIP and Tufts Face Thermal2RGB datasets for different GAN models, including pix2pix, CycleGAN, DualGAN, PCSGAN, AGGAN, AttentionGAN and TVA-GAN.

mal2RGB dataset by {79.22%, 73.63%, 78.08%, 62.20% 74.47%, 77.36%} than pix2pix, CycleGAN, DualGAN, PCSGAN, AGGAN, and AttentionGAN, respectively.

- The proposed TVA-GAN shows reduction for VGG-FaceLoss in terms of %, as reported in Table V, for WHU-IIP dataset by {22.99%, 24.16%, 19.21% , 7.83% 25.82% , 40.71%} than pix2pix, CycleGAN, DualGAN, PCSGAN, AGGAN, and AttentionGAN, respectively.
- The proposed TVA-GAN shows reduction for VGG-FaceLoss in terms of %, as reported in Table VI, for Tufts Face Thermal2RGB dataset by {44.46%, 57.60%, 56.78% , 40.37% 59.63% , 58.96%} than pix2pix, CycleGAN, DualGAN, PCSGAN, AGGAN, and AttentionGAN, respectively.

B. Qualitative Result Analysis

The qualitative result analysis between the generated images and ground truth images using the proposed TVA-GAN and different existing GAN models is shown in Fig. 6 and 7. The non-attention-based methods pix2pix, CycleGAN, DualGAN, PCSGAN, and attention-based methods AGGAN and AttentionGAN results are illustrated in Fig. 6 and 7 for Tufts Face Thermal2RGB and WHU-IIP datasets, respectively. It is visible in Fig. 6 that TVA-GAN can produce better results for more diverse datasets than the existing state-of-the-art methods. A similar observation is also made in Fig. 7 that TVA-GAN results are convincing for less diversifying dataset WHU-IIP as compared to both attention-based and non attention-based methods. Compared to the existing methods.TVA-GAN performs better than non-attention-based methods like pix2pix, CycleGAN, DualGAN, and PCSGAN have missing features due to missing attention. They do not accurately learn local as global level feature details as shown in self-attention gan [26]; on the other hand, proposed method AGGAN and AttentionGAN learning foreground and background using masking and invert masking for their method not accurately performed for fewer feature details. However, our method using recurrent inception with attention block performs better due to better segmentation using attention while translating the images.TVA-GAN is translating foreground and background information simultaneously using recurrent inception by increasing network depth and learning global and local features with fewer parameters. Generated images using TVA-GAN are more structure-preserving and close to the ground truth than results produced by other methods.

C. Impact of different losses used in TVA-GAN.

For the proposed TVA-GAN, we evaluate the impact of different losses used for training. We perform the ablation study over the Adversarial loss, Cycle loss, Cycle Synthesized loss, Synthesized loss, and Feature reconstruction losses. We can see the Qualitative comparison of various losses used in proposed method for WHU-IIP datset in Fig. 8 and for Tufts Face Thermal2RGB dataset in Fig. 9. These results are summarized as follows:

- The proposed TVA-GAN performs better than the both attention-based and non-attention-based models for thermal to visible face synthesis.
- The proposed TVA-GAN can generate more genuine visual representations using thermal face images and results in more precise details and fewer artifacts in the generated images.
- The model fails to distinguish between different person when used with only Adversarial loss on WHU-IIP dataset, and Adversarial loss with Cycle loss on Tufts Face Thermal2RGB datasets. While during training, it performs well. Hence, these two losses are not enough for generalization over the different subjects. Moreover, it is evident from the high quality generated images after combining the Adversarial loss, Cycle loss, Synthesized loss, Cycle synthesized loss and Feature reconstruction loss.

For WHU-IIP dataset, the proposed TVA-GAN leads to better SSIM, VIF and PSNR in terms of %, using combined adversarial loss,cycle loss, synthesized loss, feature reconstruction loss and combined adversarial loss,cycle loss, synthesized loss, feature reconstruction loss with Cycle synthesized loss combinations. We perceived % increment of 60.99%, 6.73%, 5.60% for SSIM, VIF and PSNR, respectively, and % reduction of 73.47% for LPIPS compared to only adversarial loss as reported in Table IX. With compared to combination of adversarial loss and cycle loss, TVA-GAN achieves the increment of 10.18%, 0.54%, 2.25% for SSIM, VIF and PSNR using combination of adversarial loss, cycle loss, synthesized loss and , while shows 37.35% of reduction for LPIPS as shown in Table IX Compared to combination of adversarial loss, cycle loss and synthesized loss, TVA-GAN gains 1.86%, 0.02%, 0.13% for SSIM, VIF and PSNR while reports a reduction of 10.34% for LPIPS by using the combination of adversarial loss, cycle loss, synthesized loss and feature

reconstruction loss as depicted in Table IX.

For Tufts Face Thermal2RGB, with combination of adversarial loss, cycle loss, cycle synthesized loss, feature reconstruction loss and synthesized loss, the proposed TVA-GAN shows improvement of 39.51%, 3.24%, 8.31% for SSIM, VIF and PSNR while shows 5.75% reduction in LPIPS as compared to the only adversarial loss as shown in Table X. With combination of adversarial loss, cycle loss, cycle synthesized loss, feature reconstruction loss and synthesized loss, the proposed TVA-GAN shows improvement over combination of adversarial and cycle loss with the gain of 42.26%, 4.04%, 9.17% for SSIM, VIF and PSNR while shows 8.08% reduction for LPIPS as reported in Table X. For Tufts Face Thermal2RGB dataset, the proposed method with all the losses also shows gain over combination of adversarial, cycle and synthesized loss by 1.63%, 0.33%, 0.19% for SSIM, VIF and PSNR while shows 1.42% reduction in LPIPS. By combining cycle synthesized loss with feature reconstruction loss, cycle loss, synthesized loss and adversarial loss, we gain 0.28%, 0.48%, 0.54% for SSIM, VIF and PSNR while shows 4.00% reduction in LPIPS for Tufts Face Thermal2RGB dataset.

D. Face verification Results for proposed TVA-GAN

For better understanding the quality of generated faces, we evaluate the generated faces using the face verification framework in this subsection. We plot the receiver operating characteristic (ROC) curves in Fig. 10 corresponding to the generated face samples using the proposed TVA-GAN with different GAN methods over the WHU-IIP and Tufts Face Thermal2RGB face datasets. We use the DeepFace [52] framework with pre-trained deep face models to calculate the distance between the generated face sample and ground truth image. We use the cosine-similarity [53] as a metric for distance calculation and use the distance as the score for the generated image samples. We use the ground truth with the corresponding generated image for the positive pairs, and for negative pairs. We use the ground-truth image with any randomly chosen generated sample from another subject. We calculate the cosine-similarity score for the positive and negative pairs and use it as a score for the ROC plot. The proposed TVA-GAN shows the gain in Fig. 10 for Face-Verification using WHU-IIP and Tufts Face Thermal2RGB datasets. For WHU-IIP face dataset, the proposed TVA-GAN shows gain of 1.177%, 9.330%, 4.182%, 5.040%, 10.259%, 15.152% compared to pix2pix, CycleGAN, DualGAN, PCS-GAN, AGGAN, and Attention-GAN. For Tufts Face Thermal2RGB dataset, the proposed method depicts improvement of 39.726%, 12.899%, 44.742%, 28.136%, 19.786%, 24.071% compared to pix2pix, CycleGAN, Dual-GAN, PCS-GAN, AGGAN and AttentionGAN.

VI. Conclusion

This paper proposes a new attention-guided generative adversarial network for thermal to visual face synthesis (TVA-GAN). The proposed model generates more realistic face images than the state-of-art methods. We design the network by including multiple losses to tackle the various problems related to image synthesis like blur, artifact generation, and semantic distortions. The losses include Adversarial loss, Cycle loss, Cycle-synthesized loss, Feature reconstruction loss and Synthesized loss.Our proposed generator network learns both local and global features accurately while transforming thermal to the visual domain . It differs in terms of only translation the foreground information as proposed in AG-GAN and AttentionGAN. It translates both the information foreground and background simultaneously without separating them. We used recurrent inception block with attention block. The proposed recurrent inception block learns the global and local features effectively, translating the images in thermal to visual domains. Recurrent inception block handles salient parts and contextual information both locally and globally by using large kernels and small kernels with more depth and fewer parameters because of the recurrent layer. While decoding occurs, attention block takes care of large receptive fields and learns more semantic contextual in formations. Attention block handles multi-stage CNN localization by crushing the feature responses in unrelated background regions progressively. The proposed model TVA-GAN is tested for the thermal to visual face synthesis problem using WHU-IIP and Tufts Face Thermal2RGB datasets. It defeats the existing state-of-the-art non-Attention-based GAN models such as pix2pix, CycleGAN, DualGAN, PCSGAN, as well as attention-based GAN models such as AGGAN and AttentionGAN. It produces more realistic faces, closer to the target image having fewer artifacts with identity preservation.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in neural information processing systems*, 2016, pp. 2802–2810.

[3] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 83–90.

[4] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 821–830.

[5] C. Peng, N. Wang, J. Li, and X. Gao, "Face sketch synthesis in the wild via deep patch representation-based probabilistic graphical model," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 172–183, 2020.

[6] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[7] R. W. Brislin, "Back-translation for cross-cultural research," *Journal of Cross-Cultural Psychology*, vol. 1, no. 3, pp. 185–216, 1970. [Online]. Available: https://doi.org/10.1177/135910457000100301

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2017.632

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[10] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[11] B. Liao and Y. Chen, "An image quality assessment algorithm based on dual-scale edge structure similarity," 10 2007, pp. 56–56.

[12] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5688–5696.

[13] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European conference on computer vision*. Springer, 2016, pp. 649–666.

[14] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[15] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?" in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4432–4441.

[16] S. Zhang, R. Ji, J. Hu, X. Lu, and X. Li, "Face sketch synthesis by multidomain adversarial learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 5, pp. 1419–1428, 2018.

[17] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[18] J. Li, P. Hao, C. Zhang, and M. Dou, "Hallucinating faces from thermal infrared images," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 465–468.

[19] J. Choi, S. Hu, S. S. Young, and L. S. Davis, "Thermal to visible face recognition," in *Sensing Technologies for Global Health, Military Medicine, Disaster Response, and Environmental Monitoring II; and Biometric Technology for Human Identification IX*, vol. 8371. International Society for Optics and Photonics, 2012, p. 83711L.

[20] C. Chen and A. Ross, "Matching thermal to visible face images using hidden factor analysis in a cascaded subspace learning framework," *Pattern Recognition Letters*, vol. 72, pp. 25–32, 2016.

[21] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 845–862, 2019.

[22] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2016, pp. 187–194.

[23] S. M. Iranmanesh, A. Dabouei, H. Kazemi, and N. M. Nasrabadi, "Deep cross polarimetric thermal-to-visible face recognition," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 166–173.

[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[25] K. K. Babu and S. R. Dubey, "Pcsgan: Perceptual cyclic-synthesized generative adversarial networks for thermal and nir to visible image transformation," *Neurocomputing*, vol. 413, pp. 41–50, 2020.

[26] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, "Unsupervised attention-guided image-to-image translation," in *Advances in Neural Information Processing Systems*, 2018, pp. 3693–3703.

[27] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-guided generative adversarial networks for unsupervised image-to-image translation," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[28] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019, pp. 7354–7363.

[29] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *ArXiv*, vol. abs/1411.1784, 2014.

[30] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in neural information processing systems*, 2016, pp. 469–477.

[31] H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *ArXiv*, vol. abs/1805.08318, 2018.

[32] A. R. Lejbølle, K. Nasrollahi, B. Krogh, and T. B. Moeslund, "Person re-identification using spatial and layer-wise attention," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1216–1231, 2020.

[33] H. Tang, H.-C. Liu, D. Xu, P. H. S. Torr, and N. Sebe, "Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks," *ArXiv*, vol. abs/1911.11897, 2019.

[34] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2417–2426.

[35] S. Ma, J. Fu, C. Wen Chen, and T. Mei, "Da-gan: Instance-level image translation by deep attention generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5657–5666.

[36] H. Tang, X. Chen, W. Wang, D. Xu, J. J. Corso, N. Sebe, and Y. Yan, "Attribute-guided sketch generation," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–7.

[37] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, "Attention-based two-stream convolutional networks for face spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 578–593, 2020.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[39] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[40] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation."

[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[42] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2813–2821.

[43] K. B. Kancharagunta and S. R. Dubey, "Csgan: Cyclic-synthesized generative adversarial networks for image-to-image transformation," *arXiv preprint arXiv:1901.03554*, 2019.

[44] Z. Wang, Z. Chen, and F. Wu, "Thermal to visible facial image translation using generative adversarial networks," *IEEE Signal Processing Letters*, vol. 25, pp. 1161–1165, 2018.

[45] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. Rao, A. Kaszowska, H. Taylor, A. Samani *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[46] S. R. Dubey, S. Chakraborty, S. K. Roy, S. Mukherjee, S. K. Singh, and B. B. Chaudhuri, "diffgrad: An optimization method for convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4500–4511, 2020.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representation*, 2015.

[48] T. Tan, S. Yin, K. Liu, and M. Wan, "On the convergence speed of amsgrad and beyond," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 2019, pp. 464–470.

[49] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 2366–2369.

[50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 586–595.

[51] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[52] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020, pp. 23–27.

[53] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *2016 4th International Conference on Cyber and IT Service Management*, 2016, pp. 1–6.