Attention-Guided Generative Adversarial Network for Explainable Thermal to Visible Face Recognition

Cunjian Chen, David Anghelone, Philippe Faure, Antitza Dantcheva

Attention-Guided Generative Adversarial Network for Explainable Thermal to Visible Face Recognition

Cunjian Chen[1,2], David Anghelone[3,4,5], Philippe Faure[4], and Antitza Dantcheva[3,5]
[1]Monash University [2]Monash Suzhou Research Institute [3]Inria [4]Thales [5]Univsersité Côte d'Azur

Abstract

Thermal to visible face image translation aims at synthesizing high-fidelity visible face images from thermal counterparts, placing emphasis on preserving the identity of the faces. While remarkable progress has been achieved related to the quality of synthetic images, as well as related to associated face matching accuracy, interpreting the generation process from thermal to visible face images remains an open challenge. Towards tackling this challenge, we present a novel generic attention-guided generative adversarial network (AG-GAN) for thermal to visible image translation. The AG-GAN framework is based on an encoder network that directly generates attention feature maps from an input thermal image in either, supervised or unsupervised fashion. A decoder network takes the attention maps and applies adaptive layer-instance normalization, in order to reconstruct the corresponding visible image. We show that solving thermal to visible image translation tasks through AG-GAN significantly improves the cross-spectral face matching accuracy, as well as inherently supports model explanation.

1. Introduction

Image-to-image translation has received increased attention due to great progress in the field of generative adversarial networks (GANs) [22, 13, 14, 16, 15]. We here are interested in thermal-to-visible image synthesis via conditional adversarial networks, which represents the task of generating photo-realistic visible face images conditioned on certain input thermal data [4, 8, 2]. This task has a wide range of applications including cross-spectral face recognition [2, 4] and face landmark detection [19], highly pertinent in defense, surveillance and public safety.

State-of-the-art thermal-to-visible image translation models have achieved reasonably high visual quality and fidelity [3]. Rakhil et al. [11] proposed a Transformers-based GAN by augmenting the network with axial-attention layers to perform simultaneous face hallucination and translation. Compared to the self-attention used in Transformers [9], the axial-attention factorizes 2D self-attention by applying 1D self-attention to height and width sequentially. Di et al. [7] presented a self-attention generative adversarial network to enhance attention-guided feature synthesis for synthesizing visible images from the polarimetric thermal inputs. However, all named works did not offer insightful *explanation* or *visualization* on which type of axial-attention or self-attention features were learned during the thermal to visible generation process. Recently, Anghelone et al. [2] utilized two separate identity and style encoders to disentangle the latent space into identity and style code representations. The associated visualization of identity code demonstrated that the identity-related structure information were well preserved during the translation. However, their work did not incorporate attention to augment the network.

In this work, we firstly introduce a generic attention-guided generative adversarial network (AG-GAN) that *encodes an input thermal image into attention feature maps*. The encoder is based on a ResNet style architecture that consists of downsampling blocks that gradually reduce the spatial size and enlarge the feature channel numbers. The decoder uses residual blocks with adaptive layer instance normalization (AdaLIN) to modulate the shape and texture change during the translation. The AdaLIN parameters are computed by applying a fully connected layer to the attention feature maps. The AG-GAN is designed to learn the attention modules, in order to guide the feature synthesis to focus on regions that are pertinent to the interests of the generator and the discriminator. Here, we consider *two types of attention* feature map learning: *supervised and unsupervised*. While the supervised attention map learns to generate the attention weights based on an auxiliary classifier, the unsupervised attention learning generates the attention weights via squeeze-excitation (SE) operation [10]. The commonality among these two approaches is that they learn the channel-based attention weights to *capture global interactions between facial contexts*. The architecture of the proposed method for the generation process is depicted in Figure 1.

The contributions of our work are summarized below.

- We design an explainable generative adversarial networks based on attention feature map learning.
- We showcase the attention maps for both, generator and discriminator.
- We compare both quantitatively and qualitatively for supervised and unsupervised attention learning.
- We offer extensive ablation studies and visualization results to validate the effectiveness of the proposed approaches from both the controlled and the uncontrolled studies.

2. Related Work

In this section, we briefly discuss existing literature on conditional adversarial networks and explainable GANs in performing general image-to-image translation.

Conditional adversarial networks [13] have so far been the de-facto model to solve image-to-image translation tasks in supervised settings. Prior works have involved notably Pix2Pix [13], aimed at learning to map a conditional input thermal image to an output visible image. The optimization step was further regularized by introducing additional constraints such as closed-set face recognition losses [21, 17] or face verification losses [4, 2], in order to preserve the identity mapping. In comparison to these, some other recent works have focused on preserving the attribute mapping by using a pre-trained attribute prediction network [12, 8]. In addition to the preservation of identity and attribute mappings, some methods [11] focused on elaborate network architecture design, incorporating the self-attention module from the Transformers [9].

Explainable GANs aim to empower image translation by transparency and interpretability. Recent works predominantly focused on the visualization and understanding of internal representations [14, 2]. Kim et al. [14] incorporated learnable attention modules into the generator and the discriminator for unsupervised image-to-image translation. Tang et al. [20] proposed an attention-guided generator to disentangle the semantic objects from the background via producing an attention mask and a content mask. The attention module was also integrated into the discriminator, focusing on attended regions only. Their proposed attention-guided generator and discriminator were used to solve unpaired image-to-image translation, which demonstrated promising results, in case that the geometric change between the source and target domain is minor.

3. Proposed Method

We propose AG-GAN, a generic attention-guided generative adversarial network designed for thermal to visible spectral translation. Our method is inspired from the U-GAT-IT work [14]. Notable differences include: (a) we re-design the entire architecture to a supervised learning framework; (b) we propose new attention feature map learning, as well as (c) we introduce a new set of loss functions such as identity loss to further constrain the mapping space. Specifically, AG-GAN encompasses three networks dedicated to the generation task, namely (i) encoder, (ii) attention and (iii) decoder.

Let $\mathcal{T}$ and $\mathcal{V}$ be the thermal and visible domains. Given an input thermal image $I_t \in \mathbb{R}^{H\times W\times C_{in}}$ and an output visible image $I_v \in \mathbb{R}^{H\times W\times C_{out}}$, the thermal-to-visible translation model can be described as:

$$\begin{array}{rccc} \Theta_{t\to v} & : & \mathcal{T} & \to & \mathcal{V} \\ & & I_t & \mapsto & \tilde{I_v} = G_v(E_t(I_t)), \end{array} \tag{1}$$

where $\Theta_{t\to v}$ consists of an encoder E_t, a decoder G_v and an auxiliary classifier $\eta_{\{\mathcal{T}\text{or}\mathcal{V}\}}$. Consequently, Equation 1 is the function producing the final output image $\tilde{I_v}$ in the visible domain. Here, H, W, C_{in} and C_{out} are the height, width, input channel number and output channel number, respectively. Let $x \in \{(I_t, I_v), (I_t, \tilde{I_v})\}$ denote a sample pair conditioned on an input thermal image I_t. Further, the discriminator D_v is adopted to determine whether x is genuine or fake. In particular, (I_t, I_v) and $(I_t, \tilde{I_v})$ denote the genuine and fake pairs, respectively.

3.1. Network Architecture

3.1.1 Generator

Encoder. Given an input thermal image I_t, we first use a 7×7 convolutional layer H_0 to transform an input image space into a high-dimensional feature space:

$$F_0 = H_0(I_t). \tag{2}$$

Here, H_0 refers to a composite function of three different operations including convolution, instance normalization and ReLU. This operation was preceded by performing a reflection padding to keep the dimensional size unchanged. Then, we apply a sequence of down-sampling operations:

$$F_i = H_i(F_{i-1}), \tag{3}$$

where F_i represents the intermediate feature maps after the i-th down-sampling operation, for all $i \in \{1, ..., K\}$ with $K \in \mathbb{N}^*$. Here, H_i is the same composite function as H_0, but with the purpose of halving the dimension and doubling the channel number. To further enhance the feature embedding, we apply a series of residual blocks H_{R_j}:

$$F_j = H_{R_j}(F_{j-1}), \tag{4}$$

where F_j, for all $j \in \{K+1, ..., M\}$ with $M > K$, denotes the intermediate feature maps after performing feature enhancement including the j-th residual block, which has two

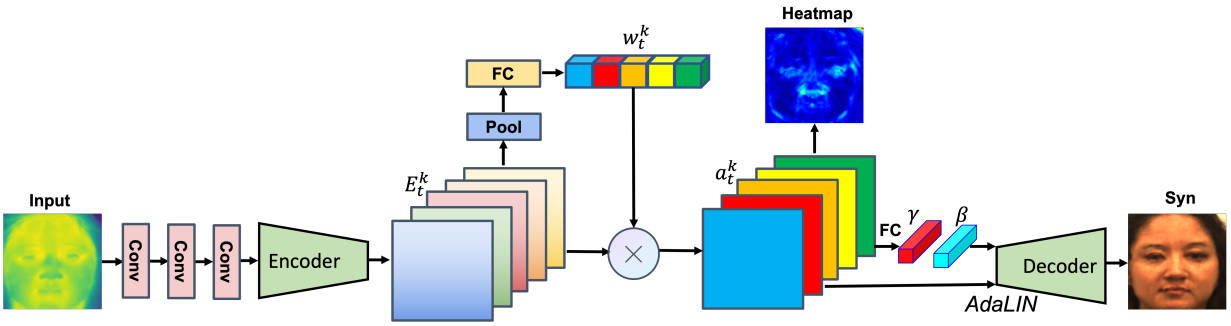

Figure 1. The architecture of the proposed AG-GAN and AG-GAN+ methods for thermal to visible image translation. The attention map in AG-GAN is generated by multiplying the learned attention weights with the feature maps obtained after encoder comprising downsampling bottlenecks. Such attention weights are obtained by inputting the GAP and GMP logits to an auxiliary classifier modulated by the CAM loss. The attention map in AG-GAN+ is generated by applying the squeeze-excitation (SE) module with no explicitly designed loss function to learn the attention weights. The decoder is formed by a series of upsampling bottlenecks coupled with AdaLIN parameters.

3×3 convolutional layers with the same output channel numbers and a skip-connection.

Attention. Given the embedding of Equations (3) and (4), we define the *encoder feature map* $E_t^k(I_t)$ as the k-th activation map from the encoder output F_M. In particular, we note $E_t^{k_{ij}}(I_t)$ as the value of activation map at (i, j). An auxiliary classifier is later introduced to learn the weight w_t^k of the k-th feature map for the thermal domain. Thus, training is driven by both, global average and global max pooling, viz. σ providing:

$$\eta_{\mathcal{T}}(I_t) = \sigma\left(\sum_k w_t^k * \sum_{i,j} E_t^{k_{ij}}(I_t)\right). \tag{5}$$

In other words, $\eta_{\mathcal{T}}(I_t)$ expresses the probability that I_t comes from the thermal domain. Finally, benefits from w_t^k provide salient thermal domain specific attention feature maps that can be illustrated as follows:

$$a_t(I_t) = w_t * E_t(I_t). \tag{6}$$

Thereby giving rise to the proposed domain translation function

$$\Theta_{t\to v}(I_t) = G_v(a_t(I_t)), \tag{7}$$

where we aim to learn the translation using neural networks. For AG-GAN+, the attention feature maps a_t are generated via the squeeze-excitation module that consists of squeeze and excitation operations. Mathematically, the squeeze operation can be described by

$$S_t = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} E_t(I_t)(i, j). \tag{8}$$

Here, H and W refer to the height and width of the encoded feature map $E_t(I_t)$ and (i, j) is the corresponding element. The channel-wise representation S_t is generated by applying the global average pooling per channel. For the excitation operation, S_t is fed into two sequential fully connected layers joined by the ReLU, which can described as follows,

$$w_t = \sigma(FC_1(ReLU(FC_2(S_t)))). \tag{9}$$

Here, w_t refers to the attention weights computed by SE. Next, we compute the resulted attention feature maps, see Eqn. 6. Note that the difference between AG-GAN and AG-GAN+ lies in the generation of attention weights w_t. The AG-GAN *supervises* the weights by the auxiliary classifier associated with the CAM loss, whereas the AG-GAN+ generates the weights by squeeze-excitation operations without employing auxiliary loss function, hence, the inception of *unsupervised* learning. The heatmap is generated by averaging the channel dimension of the attention feature map a_t.

Decoder. The decoder G_v aims at transforming a high-dimensional feature space into an output image space. It comprises of several residual blocks followed by up-sampling blocks. Here, residual blocks are instrumental in embedding features, while up-sampling convolution blocks generate target $\tilde{I}_v$ visible domain images from the associated embedded features. Inspired by U-GAT-IT [14] decoder, we further constrain the residual blocks with Adaptive Layer-Instance Normalization (AdaLIN). AdaLIN combines both advantages of Adaptive instance normalization (AdaIN) and Layer Normalization (LN) by helping the AG-GAN model to bring more flexibility in facial features generation control, with respect to shape and textures. To perform upsampling, we use the nearest neighbor strategy.

3.1.2 Discriminator

The discriminator D_v performs a binary-class classification by determining whether the given pairs (I_t, I_v) and $(I_t, \tilde{I}_v)$

are genuine or fake. This is further enhanced by constructing two different scales of PatchGAN [13] discriminators that output resulting feature maps of 6×6 and 30×30, respectively. We adopt the same attention maps used by the generator and embed them into the discriminator for both AG-GAN and AG-GAN+.

3.2. Loss Function

We utilize the pixel-wise $\mathcal{L}_1$ loss function to measure the similarity between target visible face image I_v and synthesized visible face image $\Theta_{t\to v}(I_t) = \tilde{I}_v$ at the pixel level:

$$\mathcal{L}_1 = \|I_v - \tilde{I}_v\|_1. \tag{10}$$

The objective of $\mathcal{L}_1$ loss function is to minimize the difference at the low-level features, which may not lead to the preservation of high-level features such as identity. Therefore, we utilize a pre-trained ArcFace recognition network [6] to extract the face feature embedding, measure the cosine similarity and compute the identity loss function:

$$\mathcal{L}_{ID} = 1 - < \phi_R(I_t), \phi_R(\tilde{I}_v) > . \tag{11}$$

Here, $\phi_R(I_t)$ and $\phi_R(\tilde{I}_v)$ denote the normalized feature embedding. Further, the CAM loss results from the auxiliary classifier can be described as:

$$\mathcal{L}_{CAM}^{\Theta_{t\to v}} = -(\mathbb{E}_{x\sim\mathcal{T}}[log(\eta_{\mathcal{T}}(x))] + \mathbb{E}_{x\sim\mathcal{V}}[1 - log(\eta_{\mathcal{T}}(x))]) \tag{12}$$

$$\mathcal{L}_{CAM}^{D_v} = \mathbb{E}_{x\sim\mathcal{V}}[\eta_{\mathcal{V}}(x)^2] + \mathbb{E}_{x\sim\mathcal{T}}[(1-\eta_{\mathcal{V}}(G_v(x)))^2]. \tag{13}$$

Note that for AG-GAN+, no auxiliary classifier was used to learn the attention weights, thus no CAM loss was required. We also introduce the perceptual loss to enhance the image quality:

$$\mathcal{L}_P = \|\phi_P(I_v) - \phi_P(\tilde{I}_v)\|_1. \tag{14}$$

Here, ϕ_P denotes the perceptual network constructed from the VGG-19. The overall loss function is the combination of aforementioned loss functions, along with the adversarial loss function.

3.3. Implementation Details

We use LSGAN for training the AG-GAN by setting weight of $\mathcal{L}_1$, $\mathcal{L}_{ID}$, and $\mathcal{L}_{CAM}^{\Theta_{t\to v}}$ to be 100, 1, and 1000, respectively. The default number of epochs used in our training was 200. Images were first scaled to size 286×286, and then randomly cropped to a size of 256×256. A batch size of 1 with the Adam optimizer was used. The perceptual loss $\mathcal{L}_P$ was used only in SpeakingFaces experiment. During the inference process, we experimented with models generated by different epochs and select epoch 90 as our final model.

4. Experimental Results

To verify the effectiveness of the proposed method, experiments were conducted on the controlled ARL-VTF [18] dataset, as well as the uncontrolled SpeakingFaces [1] dataset.

4.1. Evaluation on ARL Dataset

The ARL-VTF dataset [18] is considered as the largest collection of paired thermal and visible face images. Following the established evaluation protocol, 295 subjects were used for training and 100 subjects were used for testing. Specifically, the setting of gallery (G_VB0-) and probe (P_TB0-) subjects without glasses was chosen. The face images were pre-processed by aligning and cropping based on the manually annotated landmarks to minimize variations unrelated to the identity (see Figure 6).

Table 1. Comparison of proposed method with GAN-based face matching methods on ARL-VTF dataset.

Method	AUC↑	EER↓	PSNR↑	SSIM↑
U-GAT-IT [14]	89.58	17.84	14.49	0.65
AttentionGAN [20]	93.35	13.41	14.34	0.64
LG-GAN [2]	94.26	12.99	15.77	0.61
SG-GAN [4]	98.52	5.07	17.27	0.71
Axial-GAN [11]	99.05	4.98	-	-
AG-GAN	98.74	5.56	17.39	0.70
AG-GAN+	**99.26**	**4.30**	**17.58**	**0.72**

Table 1 shows quantitative comparisons between (a) proposed methods: AG-GAN and AG-GAN+ and (b) state-of-the-art methods: U-GAT-IT [14], AttentionGAN [20], LG-GAN [2], SG-GAN [4] and Axial-GAN [11]. As showcased in Table 1, AG-GAN+ achieves best performance on all evaluation metrics. Note that U-GAT-IT and AttentionGAN were designed for unsupervised image-to-image translation. The remaining methods are applicable to supervised image-to-image translation. For a fair comparison with Axial-GAN, we adopt the face matching results from [11]. The maximum AUC is achieved by AG-GAN+. Here, the AUC and EER are used to measure the face matching accuracy, while PSNR and SSIM represent the image quality. We observe that PSNR and SSIM are positively correlated with AUC and EER. Additionally, we show corresponding ROC curves in Figure 2. The proposed methods achieve significantly higher true match rates across different false match rates. Further, a qualitative comparison between the selected methods is shown in Figure 3. Notably, the proposed methods AG-GAN and AG-GAN+ are able to accurately reconstruct visible images from thermal inputs, while preserving well identity and incorporating finer details. In contrast, U-GAT-IT, AttentionGAN and LG-GAN appear to include more artifacts.

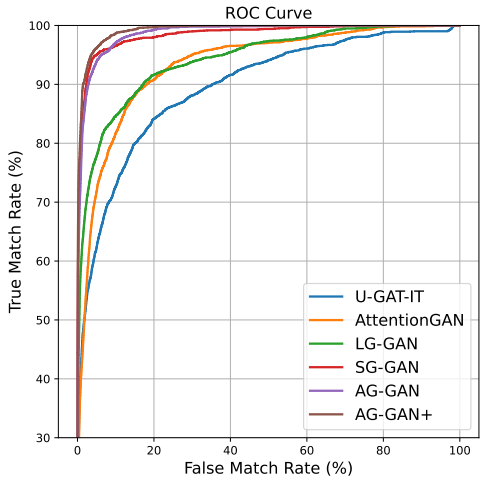

Figure 2. ROC results of proposed algorithms and existing works on ARL-VTF dataset.

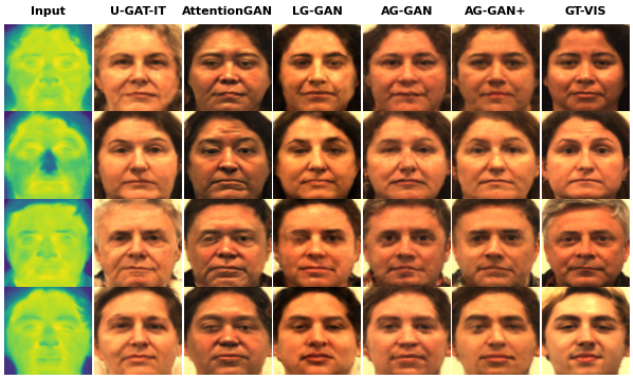

Figure 3. Comparison of qualitative results of proposed algorithms with existing works on ARL-VTF dataset.

4.2. Evaluation on SpeakingFaces

The SpeakingFaces dataset [1] consists of 142 subjects, where 100 subjects were used for training and the remaining subjects were used for testing. Following the established protocol used in [1], 5,400 and 2,268 thermal-visible image pairs were utilized for training and testing, respectively. Each subject was captured under 9 different poses, making it suitable for evaluating thermal-to-visible image translation under pose variations.

Table 2 shows quantitative comparisons between (a) proposed methods: AG-GAN and AG-GAN+ and (b) state-of-the-art methods: U-GAT-IT [14], CUT [1], AttentionGAN [20], Pix2Pix [13], CycleGAN [1], SG-GAN [4], Axial-GAN [11][1]. As demonstrated in Table 2, the proposed AG-GAN and AG-GAN+ methods achieve the best overall results for evaluating thermal-to-visible image translation under pose variation. This is corroborated by the use of face verification and image quality metrics. Further, we compute the ROC curves for selected algorithms in Figure 5. In addition, a qualitative comparison between the methods is shown in Figure 4. Both proposed methods are able to synthesize high-fidelity visible face images under pose variations, while preserving well the identity.

[1]Axial-GAN has been reproduced on the high-resolution thermal to visible image translation task. We make sure the results from the reproduced model is similar to the original results.

Table 2. Comparison of proposed method with GAN-based face matching methods on SpeakingFaces dataset.

Method	AUC↑	EER↓	PSNR↑	SSIM↑
U-GAT-IT [14]	83.0	24.48	19.05	0.71
CUT [1]	83.62	23.59	20.51	0.67
AttentionGAN [20]	84.69	22.92	19.21	0.71
Pix2Pix [13]	86.82	20.99	20.29	0.72
CycleGAN [1]	86.97	20.70	20.34	0.67
SG-GAN [4]	88.41	19.23	20.33	0.72
Axial-GAN [11]	89.51	17.90	**21.15**	0.69
AG-GAN	89.86	17.68	20.82	0.74
AG-GAN+	**90.53**	**17.20**	21.01	**0.75**

4.3. Ablation Study

We conduct an ablation study to understand the effectiveness of AG-GAN with respect to the GCAM loss $\mathcal{L}_{CAM}^{\Theta_{t\to v}}$, DCAM loss $\mathcal{L}_{CAM}^{D_v}$ and identity loss $\mathcal{L}_{ID}$.

Impact of CAM loss. As seen in Table 3, the use of CAM loss can increase the face matching accuracy by pushing the generator to focus on salient facial regions. Here, "w/o DCAM" and "w/o GCAM" refers to the settings, where no CAM loss is applied to the discriminator and generator, respectively. "w/o GDCAM" refers to the setting where no CAM loss is applied to both, generator and discriminator. However, we do not observe a correlation between face matching accuracy and perceived image quality when analyzing the CAM loss.

Table 3. Ablation study on the impact of CAM loss with ARL-VTF dataset.

Method	AUC↑	EER↓	PSNR↑	SSIM↑
AG-GAN	**98.74**	**5.56**	17.39	0.70
w/o DCAM	97.19	7.09	17.00	0.69
w/o GCAM	97.93	6.30	17.67	0.70
w/o GDCAM	97.74	7.29	**17.81**	**0.71**

Impact of Identity loss. We investigate the use of different face recognition networks to extract feature embedding and compute identity loss. As seen in Table 4, using ArcFace [6] to derive the feature embedding for identity loss computation results in higher performance than MobileFaceNet [5], as well as the framework without adopting the identity loss. Admittedly, applying face recognition loss does not always result in improved matching performance. Therefore, it is

Figure 4. Comparison of qualitative results of proposed algorithms with existing works on SpeakingFaces dataset.

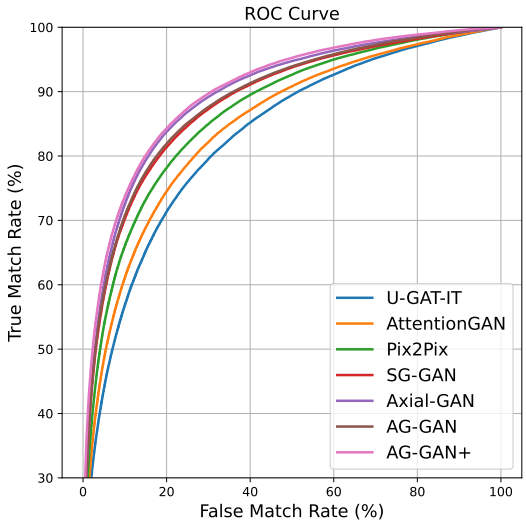

Figure 5. ROC results of proposed algorithms and existing works on SpeakingFaces dataset.

critical to choose a suitable identity loss to ensure maximum identity similarity between target visible and synthesized visible images.

Table 4. Ablation study on the impact of identity loss with ARL-VTF dataset using AG-GAN.

Method	AUC↑	EER↓	PSNR↑	SSIM↑
w/o ID	97.66	7.78	17.02	0.69
w/ MobileFaceNet	97.59	7.57	16.94	0.69
w/ ArcFace	**98.74**	**5.56**	**17.39**	**0.70**

4.4. Results on Attention Maps

For interpretation of the generation process from input thermal to output visible images, we visualize the attention maps embedded in the generator. As seen in Figure 6, features are generally activated around the salient facial regions including nose, mouth and eyes. Other notable regions such as hair are also likely to be activated. However, it is worth pointing out that the skin regions are not activated in the resulting attention maps. This clearly indicates that the generator network tends to focus on salient facial features that are discriminative across subjects. The skin region, on the other hand, appears to share more similarity in texture, thereby not being accentuated by the generator. In contrast, the discriminator attention maps focus on distinguishing the skin region difference between the synthesized and ground-truth visible face images. This could be explained by the fact that features from salient facial regions are well synthesized, hence considering to be genuine by the discriminator. Thus, the generator and discriminator are unlikely to compete against each other on these regions to distinguish between genuine and fake images.

We show that the attention maps entail highly similar representations throughout the entire generation process. Figure 7 reveals the attention maps produced at different test epochs for a single subject. It becomes evident that attention maps at different stages demonstrate highly consistent representations in salient facial regions including eyes, nose and mouth. This clearly validates that the visual attention is consistent across the entire training process. Note that with the increase of epochs, more visually pleasing images can be obtained.

Observation of Pose. We observe that such attention maps are also effective in interpreting the thermal-to-visible image translation in uncontrolled conditions, e.g., pose variation. Figure 8 presents the heatmap visualization results for a subject captured under 9 different pose positions. Apparently, the heatmap interpretation is also consistent with the conclusion drawn from Figure 6.

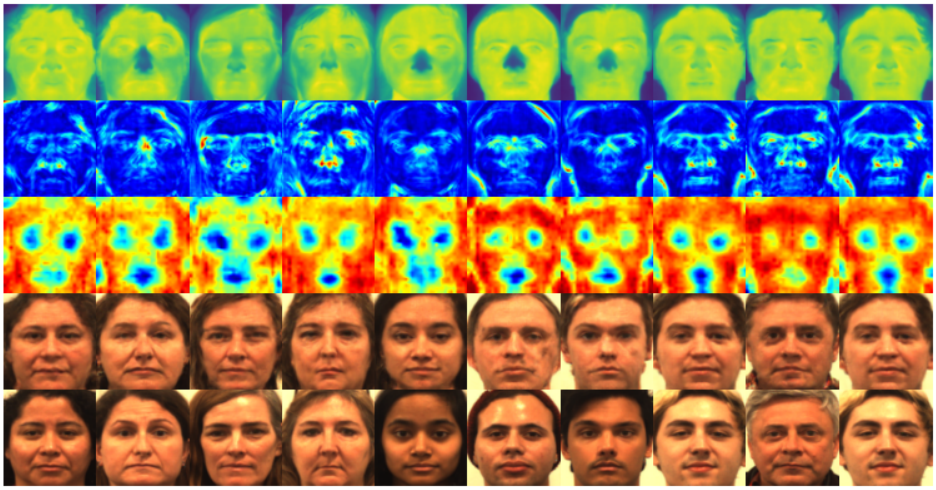

Figure 6. **Transparency and Interpretability.** Examples of attention maps produced by the generator and discriminator on ARL-VTF dataset using AG-GAN. The images from top to bottom rows are: thermal, generator attention map, discriminator attention map, synthesized visible and ground-truth visible face images.

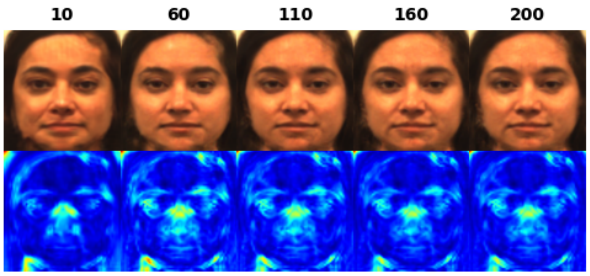

Figure 7. **Robustness and Consistency.** Examples of attention maps produced by the generator at individual test epochs on ARL-VTF dataset using AG-GAN. The images in the top and bottom rows are synthesized visible images and their corresponding attention maps.

4.5. Supervised vs. Unsupervised Attention

The attention weights from the proposed AG-GAN method were learned by the CAM loss, while the attention weights from AG-GAN+ were learned by the squeeze-excitation operation. Here, the supervised and unsupervised learning were defined based on whether the attention weights are directly learned by a loss function. Figure 9 shows the heatmaps generated by the squeeze-excitation. Compared to Figure 6, where the attention maps are more localized, the AG-GAN+ generates the attention maps that are more globally distributed. This is challenging for the model explainability, as there are no consistent patterns observed. However, the attention maps computed by the generator still clearly reveal the structure information.

5. Conclusions

In this paper, we propose a general-purpose attention-guided generative adversarial model for explainable thermal-to-visible image translation. Attention maps are extracted from the encoder in both, supervised and unsupervised manner and are fed into the AdaLIN-based decoder. By visualizing the learned attention maps, we show that AG-GAN is capable of interpreting thermal-to-visible image translation. Additionally, we demonstrate that the proposed methods achieve competitive cross-spectral face matching performances. Future work will explore the spatial attention and the self-attention from Transformers, in order to further improve the performance.

References

[1] M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, Y. Khassanov, M. Lewis, and H. A. Varol. Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams. *Sensors*, 21(10):3465, 2021.

[2] D. Anghelone, C. Chen, P. Faure, A. Ross, and A. Dantcheva. Explainable thermal to visible face recognition using latent-guided generative adversarial network. In *Proc. of FG*, 2021.

[3] D. Anghelone, C. Chen, A. Ross, and A. Dantcheval. Beyond the visible: A survey on cross-spectral face recognition. In *arXiv Preprint*, 2022.

[4] C. Chen and A. Ross. Matching thermal to visible face images using a semantic-guided generative adversarial network. In *Proc. of FG*, 2019.

[5] S. Chen et al. Mobilefacenets: Efficient CNNs for accurate real-time face verification on mobile devices. In *Proc. of CCBR*, 2018.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proc. of CVPR*, 2019.

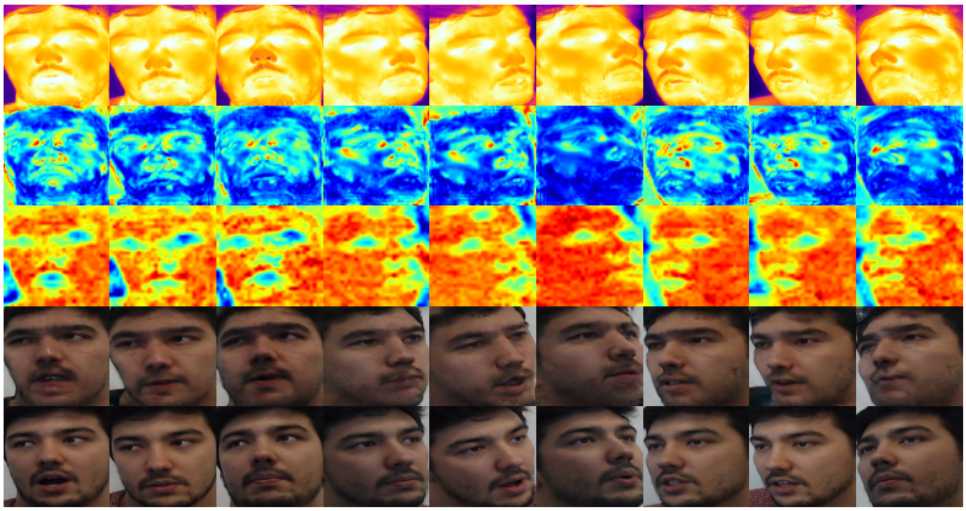

Figure 8. **Transparency and Interpretability.** Examples of attention maps from 9 different poses of the same subject produced by the generator and discriminator on SpeakingFace dataset using AG-GAN. The images from top to bottom rows are: thermal, generator attention map, discriminator attention map, synthesized visible and ground-truth visible face images.

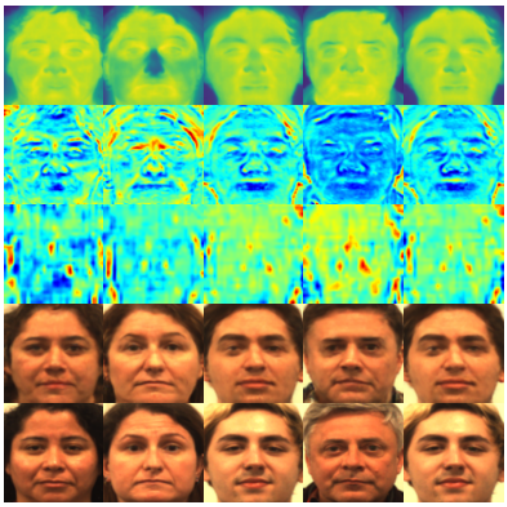

Figure 9. Examples of attention maps produced by unsupervised attention learning using AG-GAN+, ordered by thermal, generator and discriminator attentions, synthesized and ground-truth visible from ARL-VTF dataset.

[7] X. Di, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel. Polarimetric thermal to visible face verification via self-attention guided synthesis. In *Proc. of ICB*, 2019.

[8] X. Di, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel. Multi-scale thermal to visible face verification via attribute guided synthesis. *IEEE Trans. on BIOM*, 3(2):266–280, 2021.

[9] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[10] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proc. of CVPR*, 2018.

[11] R. Immidisetti, S. Hu, and V. M. Patel. Simultaneous face hallucination and translation for thermal to visible face verification using axial-gan. In *Proc. of IJCB*, 2021.

[12] S. M. Iranmanesh and N. M. Nasrabadi. Attribute-guided deep polarimetric thermal-to-visible face recognition. In *Proc. of ICB*, 2019.

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

[14] J. Kim, M. Kim, H. Kang, and K. Lee. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *Proc. of ICLR*, 2020.

[15] K. Mei, Y. Mei, and V. M. Patel. Thermal to visible image synthesis under atmospheric turbulence. In *arXiv Preprint*, 2022.

[16] Y. Mei, P. Guo, and V. M. Patel. Escaping data scarcity for high-resolution heterogeneous face hallucination. In *Proc. of CVPR*, 2022.

[17] N. Peri, J. Gleason, C. D. Castillo, T. Bourlai, V. M. Patel, and R. Chellappa. A synthesis-based approach for thermal-to-visible face verification. In *Proc. of FG*, 2021.

[18] D. Poster et al. A large-scale, time-synchronized visible and thermal face dataset. In *Proc. of WACV*, 2021.

[19] D. Poster, S. Hu, N. J. Short, B. S. Riggan, and N. M. Nasrabadi. Visible-to-thermal transfer learning for facial landmark detection. *IEEE Access*, 9:52759–52772, 2021.

[20] H. Tang, H. Liu, D. Xu, P. H. Torr, and N. Sebe. Attention-GAN: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Trans. on NNLS*, 2021.

[21] T. Zhang, A. Wiliem, S. Yang, and B. C. Lovell. TV-GAN: generative adversarial network based thermal to visible face recognition. In *Proc. of ICB*, 2018.

[22] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of ICCV*, 2017.