ANYRES: Generating High-Resolution visible-face images from Low-Resolution thermal-face images

David Anghelone, Sarah Lannes, Antitza Dantcheva

ANYRES: GENERATING HIGH-RESOLUTION VISIBLE-FACE IMAGES FROM LOW-RESOLUTION THERMAL-FACE IMAGES

David Anghelone[1,2,3], *Sarah Lannes*[2], *Antitza Dantcheva*[1,3]
[1]Inria [2]Thales [3]Université Côte d'Azur

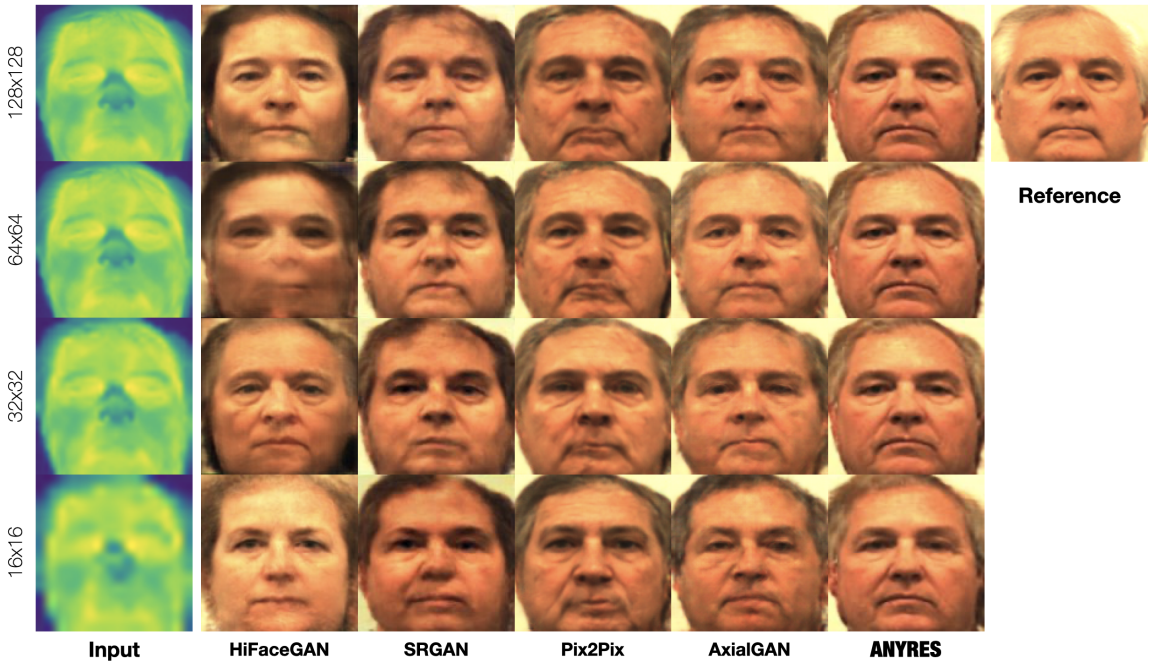

Fig. 1. **Qualitative results** of HiFaceGAN, SRGAN, Pix2Pix, AxialGAN and the proposed ANYRES on the ARL-VTF dataset. We decrease resolution in each row (re-scaled to 128×128). While previous methods are impaired to super resolve facial images for a given resolution by using one specific network for each resolution, our proposed ANYRES achieves a balance between realism and fidelity across resolutions with solely one unified network.

ABSTRACT

Cross-spectral Face Recognition (CFR) aims to compare facial images across different modalities, *i.e.,* the visible and thermal spectra. CFR is more challenging than traditional face recognition (FR) due to the profound modality gap in-between spectra. As related applications range from night-vision FR to robust presentation attacks detection, acquisition involves capturing images at varying distances, represented by different image resolutions. Prior approaches have addressed CFR by considering a fixed resolution, necessitating that a subject stands at a precise distance from a given sensor during acquisition, which constitutes an impracticable scenario in real-life. Towards loosening this constraint, we propose ANYRES, a unified model endowed with the ability to handle a *wide range of input resolutions*. ANYRES generates high resolution visible images from low resolution thermal images, placing emphasis on *maintaining the cross-spectral identity*. We demonstrate the effectiveness of the method and present extensive FR experiments on multi-spectral paired face datasets.

Index Terms— Biometrics, Cross-spectral Face Recognition, Super Resolution, Generative Models.

1. INTRODUCTION

Thermal sensors play a crucial role in detecting and recognizing humans in surveillance settings, specifically in the context of long-range distance acquisition or under adverse lighting conditions (low-light or night-time environments). However, thermal imaging does not provide detailed rendering of faces, hindering related FR systems. Therefore, generating visible-spectrum face images of High-Resolution (HR) based on associated thermal-spectrum face images of Low spatial Resolution (LR) is of particular pertinence in designing operational CFR systems [1], where for example a visible face image is compared to a face image acquired beyond the visible spectrum. Such generating process is referred to as Super Resolution (SR) or Hallucination, aiming to produce HR images based on single or sequential LR images. All existing

solutions allow for SR from a *fixed input resolution* [2], making them completely impractical in real-life scenarios.

Motivated by the above, we here address the task of comparing thermal face images of any (low) resolution against a gallery of HR visible face images by designing a unique model handling dual computer vision tasks, i.e. *super resolution* and *domain translation*, streamlined to be more adaptive to ensure faithful cross-spectral identity preservation. In particular, we propose ANYRES, a novel model that allows for simultaneous face SR, as well as thermal-to-visible spectrum translation. We have place emphasis on ANYRES being robust to any LR thermal inputs, while preserving the identity. Benefits from the simultaneous process are instrumental in avoiding accumulated errors and artifacts. ANYRES is equipped to bridge simultaneously the modality gap, as well as the resolution gap. In particular, a blurry, thermal LR face image is transformed into a sharp, realistic, HR visible face image. The designed network presents the advantage of preserving consistent biometric features across both, the LR/HR space, as well as the thermal/visible spectrum, allowing for comparison of super resolved images and a gallery of visible images, using off-the-shelf FR algorithms. Furthermore, the proposed algorithm is suitable to real world scenarios as during operational applications humans are randomly situated away from the camera and can therefore depict multi-scale LR thermal face images (which depends of the acquisition distance). Unlike the state-of-the-art, where the resolution is generally fixed as input, ANYRES is to the best of our knowledge the first framework to operate at any input resolution ranging from LR to HR.

The main contributions of this work include the following.

- We propose a novel supervised learning framework for CFR that performs simultaneously both, *domain translation* and *super-resolution*. Specific loss functions have been introduced, in order to enhance both, image quality as well as biometric feature preservation.
- We empower the network by learning to process a range of resolutions as inputs, while previous methods enabled only fixed input resolution. Our mechanism is based on a resolution-*inter-dependency*, (i) taking advantage of pyramidal architecture, as features are perceived with multi-scale analysis and (ii) gating spectral encoded features with decoded super resolved features.
- We achieve state-of-the-art performance on four benchmark multi-spectral face datasets, with respect to *visual quality*, as well as *face recognition* scores.

2. RELATED WORK

Large resolution discrepancy between visible and thermal sensors induces paired visible-thermal face datasets with images having significant resolution-gap [1]. Although existing CFR methods [3, 4] are based on GANs to simulate artificial visible-like facial images from thermal face images, they have not considered the aspect of changing resolution. They predominantly focused on conditional-GAN, where multi-spectral paired facial samples were used in a supervised learning. Consequently, Pix2Pix designed with UNet-like encoder-decoder architecture was adopted for its ability to learn conditional mapping from one domain to another. Further optimization was introduced to constrain *perceptual*-rendering [5], *identity*-preservation [6], and *semantic*-attribute guidance [7], between the synthesized visible face images and the target visible face images. Opening the work on reliable CFR system in unconstrained environment, Immidisetti et al. [2] proposed the first study dealing with resolution for long-range surveillance system. Their work entitled AxialGAN attempt to perform CFR when humans are distant away from the camera. AxialGAN addresses simultaneously spectrum translation from thermal-to-visible and face hallucination, but restricted to process a fixed input resolution. The proposed GAN framework designed an axial-attention layer to capture long-range dependencies, incorporated into both generator and discriminator networks.

3. PROPOSED METHOD

We propose ANYRES, a GAN streamlined to address simultaneously both tasks, *domain translation* and *super resolution*, while preserving identity. In particular, ANYRES tackles the problem of matching any LR thermal face image against HR visible face images by (i) learning an end-to-end mapping between the thermal spectrum and the visible spectrum, and (ii) learning to handle input of any resolution.

3.1. Problem Formulation

We here consider the HR space, with cardinality $m \times n$, incorporating a visible domain $\mathcal{V}$ with visible face images $x_{vis} \in \mathbb{R}^{m \times n}$, and a thermal domain $\mathcal{T}$ with thermal face images $x_{thm} \in \mathbb{R}^{m \times n}$.

Domain translation phase In the domain translation phase, image-to-image translation is performed by learning an end-to-end non-linear mapping, denoted as $\Theta_{t \to v}$, between the thermal spectrum and the visible spectrum. This is formalized as follows:

$$\begin{array}{ccccc} \Theta_{t \to v} & : & \mathcal{T} & \to & \mathcal{V} \\ & & x_{thm} & \mapsto & x_{vis}^{synthetic}. \end{array} \quad (1)$$

$\Theta_{t \to v}$ represents the function that synthesizes the corresponding thermal face images into a realistic synthetic visible face images $x_{vis}^{synthetic}$ in the HR space.

Super Resolution phase. Given the embedding of Equation (1), the network encapsulates the SR scalability as a simultaneous task. Therefore, we aim to learn a conditional generation function, where a thermal LR facial image $x_{thm}^{LR} \in$

$\mathbb{R}^{\frac{m}{r}\times\frac{n}{r}}$ is also enhanced to the HR scale, providing a synthetic visible image $x_{vis}^{SR} \in \mathbb{R}^{m\times n}$ up-scaled by a $\times r > 0$ scale factor, via:

$$x_{vis}^{SR} = \Theta_{t\to v}(x_{thm}^{LR}). \tag{2}$$

As elaborated above, thermal-to-visible FR based on GAN-synthesis, with the objective of being robust to any LR thermal inputs, aims to learn a unified function that, when applied to *any*-LR thermal image x_{thm}^{LR}, yields a higher-resolution super resolved (SR) visible image $x_{vis}^{SR} \in \mathbb{R}^{m\times n}$ with rich semantic and identity information. In this context, the contribution of ANYRES is the simultaneous learning of global interaction between both *domain translation* and *resolution* scalability through the enrichment of Equation (1) by Equation (2). To be specific, for all scale factor $0 < r \leqslant m$, the method $\Theta_{t\to v}$ is designed to learn neural networks by considering (x_{thm}, x_{vis})-paired facial images and minimizing specific loss functions (supervised setting).

3.2. Baseline Model

Towards learning how to process any resolution as input, without having to estimate of said resolution, ANYRES is based on a U-shape pyramidal architecture. It relies naturally on a multi-scale analysis. The overall architecture is illustrated in Figure 2.

We model our function $\Theta_{t\to v}$ using a U-net architecture. The generator consists of an encoder-decoder structure with skip connections between domain specific encoder and decoder. Considering the larger discrepancy between the images resulted from LR and HR spaces, we introduce *Squeeze-and-Excitation* [8] (SE) blocks, which play the role of gate modulator after each skip connection. Such strategy enables channel-wise relationships and brings a flexible control for balancing encoded features with decoded super resolved features.

3.2.1. Generator

Cross-Resolution Interaction. During training time, the network is fed simultaneously with batches of a wide range r-scale factors of (low) resolution thermal images. In the case of a fixed r, the model is able to super resolve images from $\frac{m}{r}\times\frac{n}{r}$ to $m\times n$ scale of space (i.e. fixed LR input unlike any LR input). In what follows, we refer to a model trained with one scale factor as *mono-resolution*, whereas a model trained with several scale factor is denoted *multi-resolution*.

Encoder. The encoder extracts multi-resolution features in parallel and fuses them repeatedly during the learning stage, in order to generate high-quality SR-*representations* with rich semantic/identity information.

Given a LR thermal input image x_{thm}^{LR}, we first use a layer H_0 to transform a LR input image space into a high-dimensional feature space:

$$F_0 = H_0(x_{thm}^{LR}). \tag{3}$$

Here, H_0 refers to a composite function of two successive *Convolution-BatchNormalization-ReLU* layers. Then, we apply a sequence of operations:

$$F_i = H_i(\text{Pool}(F_{i-1})), \tag{4}$$

where F_i represents the intermediate encoded feature maps after the i-th operation, for all $i \in [1, K]$ with $K \in \mathbb{N}^*$. Here, H_i is the same composite function defined in Equation (3), and Pool denotes a max pooling operation where the most prominent features of the prior feature map are preserved.

Decoder The decoder aims at transforming a high-dimensional feature space into a SR output image in the visible spectrum. Hence, the generative task towards the super resolved images is started from the deep level (U bottleneck),

$$G_K = H_K(SE(C(F_{K-1}, S_K(F_K)))). \tag{5}$$

Then sequentially incremented, for all $i \in [1, K-1]$, by

$$G_i = H_i(SE(C(F_{i-1}, S_i(G_{i+1})))), \tag{6}$$

ended by the generation of the SR image x_{vis}^{SR} through *Convolution-Tanh* layers

$$G_0 = x_{vis}^{SR}. \tag{7}$$

While S refers to upsampling operation by factor 2 followed by *Convolution-BatchNormalization-ReLU* layers, C concatenates all channels from the skip connection F_{i-1} with the up-sampled S_i layers. Finally, G_i represents the decoded intermediate feature maps after the i-th operation preceded by Squeeze and Excitation SE.

3.2.2. Discriminators

In an adversarial learning, ANYRES is complemented by global and local discriminators, named $\textbf{Dis}_{\text{global}}$ and $\textbf{Dis}_{\text{local}}$ respectively, further depicted in Figure 3. The former helps the generator to synthesize photo-realistic HR image, whereas the latter is focused on subtle facial details and benefits from local inherent attention to capture faithful biometric features during the generation.

Global Discriminator. We adopt the multi-scale discriminator, which enables generation of realistic images with refined details. Hence, $\textbf{Dis}_{\text{global}}$ is responsible of performing a binary-classification by distinguishing super resolved image x_{vis}^{SR} from real image x_{vis}.

Local Discriminator. To synthesize biometric-realistic semantic content, we focus on discriminant areas relevant for identification. Such regions are represented by the same cropping area (see Figure 3) between the images x_{vis} and x_{vis}^{SR}, respectively named $x_{vis-ROI,i}$ or $x_{vis-ROI,i}^{SR}$, with $i \in [0, 4]$. Thus, the local discriminator $\textbf{Dis}_{\text{local}}$ extends the design to independent discriminators L_i, paying attention to every single facial fine details and benefit from local inherent attention to capture faithful biometric features during the generation.

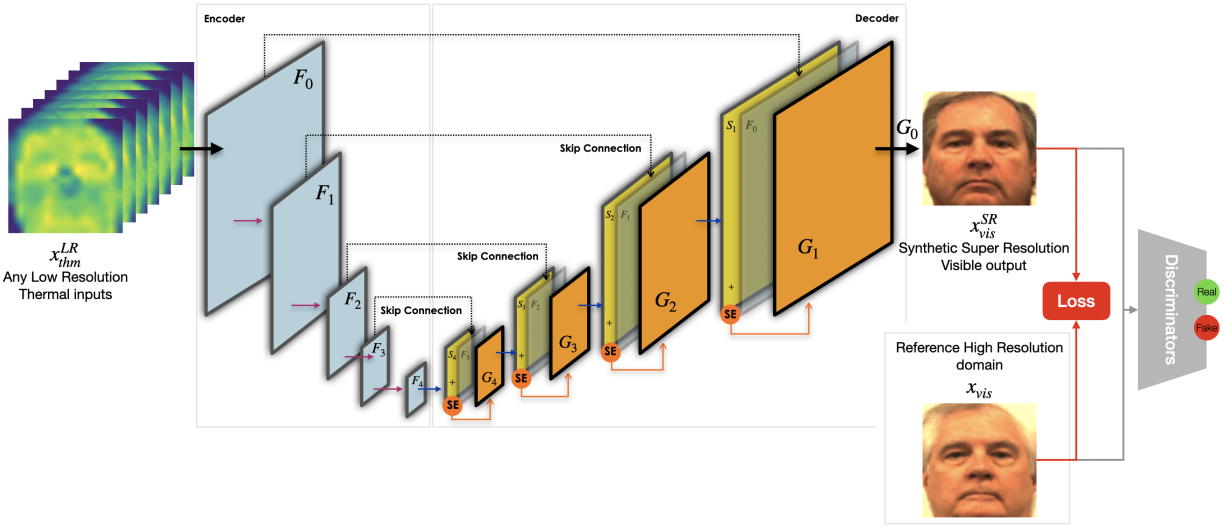

Fig. 2. **Training of ANYRES.** The generator accepts any (low)-resolution thermal face x_{thm}^{LR} as input. It comprises an encoder-decoder bridged by skip connections and gated by Squeeze and Excitation (SE) blocks, which play the role of gate modulator and enable resolution-wise relationships towards bringing a flexible control for balancing encoded features with decoded super resolved features. The discriminators are aimed at distinguishing real images x_{vis} from generated synthetic ones x_{vis}^{SR}.

3.3. Loss Functions

The learning process of ANYRES is driven by an efficient combination of objective functions that pave the way to control the synthesis process at both *pixels* and *features* levels.

Adversarial loss. Images generated through Equation (1) must be realistic. Therefore, the objective of the generator is to maximize the probability of the discriminators making incorrect decisions. The objective of the discriminators, on the other hand, is to maximize the probability of making a correct decision, i.e., to effectively distinguish between real and synthesized images. The global $\mathcal{L}_{GAN}^{Global}$ and local $\mathcal{L}_{GAN}^{Local}$ loss functions are part of the adversarial training and defined as follows:

$$\mathcal{L}_{GAN} = \mathcal{L}_{GAN}^{Global} + \mathcal{L}_{GAN}^{Local}. \tag{8}$$

Conditional loss. Imposing a condition on the spectral distribution is essential for generating images within the target spectrum. The conditional loss (known as L1 loss) is defined as follows:

$$\mathcal{L}_{cond} = \mathop{\mathbb{E}}_{x_{vis}; x_{vis}^{SR} \sim p_{\mathcal{V}}} \|x_{vis}^{SR} - x_{vis}\|_1. \tag{9}$$

Perceptual loss. The perceptual loss $\mathcal{L}_P$ affects the perceptive rendering of the image (ensuring they are representing faces) by measuring the high-level semantic difference between synthesized and target face images. It reduce artefacts and enables the reproduction of realistic details. $\mathcal{L}_P$ is defined as follows:

$$\mathcal{L}_P = \mathop{\mathbb{E}}_{x_{vis}; x_{vis}^{SR} \sim p_{\mathcal{V}}} \|\phi_P(x_{vis}^{SR}) - \phi_P(x_{vis})\|_1, \tag{10}$$

where, ϕ_P represents features extracted by VGG-19, pre-trained on ImageNet.

Identity loss. The identity loss $\mathcal{L}_I$ preserves the identity of the facial input and relies on a pre-trained ArcFace [9] recognition network to extract facial features embedding. Then, cosine similarity measure provides the identity loss function:

$$\mathcal{L}_I = \mathop{\mathbb{E}}_{x_{vis}; x_{vis}^{SR} \sim p_{\mathcal{V}}} [1 - < \phi_I(x_{vis}), \phi_I(x_{vis}^{SR}) >], \tag{11}$$

where, ϕ_I denotes the features extracted from Arcface.

Attribute loss. The attribute loss $\mathcal{L}_A$ prevents attribute shift during spectrum translation. While age brings apparent information, gender relies on identity. Therefore, apparent age loss $\mathcal{L}_A^{Age}$ and gender loss $\mathcal{L}_A^{Gender}$ are defined as follows:

$$\mathcal{L}_A^{Age} = \mathop{\mathbb{E}}_{x_{vis}; x_{vis}^{SR} \sim p_{\mathcal{V}}} \|\phi_{Age}(x_{vis}^{SR}) - \phi_{Age}(x_{vis})\|_1, \tag{12}$$

$$\mathcal{L}_A^{Gender} = \mathop{\mathbb{E}}_{x_{vis}; x_{vis}^{SR} \sim p_{\mathcal{V}}} \|\phi_{Gender}(x_{vis}^{SR}) - \phi_{Gender}(x_{vis})\|_1, \tag{13}$$

where, ϕ_{Age} and ϕ_{Gender} are pre-trained models based on DeepFace facial attribute framework analysis [10]. Then, the attribute loss is denoted as follow: $\mathcal{L}_A = \mathcal{L}_A^{Age} + \mathcal{L}_A^{Gender}$.

Finally, all loss functions combined together bring realism during spectral translation and avoid blurriness introduced by any low scale of resolution from thermal image inputs. ANYRES relies on the combination of the aforementioned loss functions.

3.4. Implementation Details

ANYRES is implemented in PyTorch and uses Adam optimizer with an initial learning rate of 0.0002, and $\beta_1 = 0.5$, $\beta_2 = 0.999$. For all experiments, the batch size and the default number of epochs used are set to 4 and 100, respectively.

Images are first aligned with eyes, nose and mouth key points by following the protocol expressed in [11] and scaled

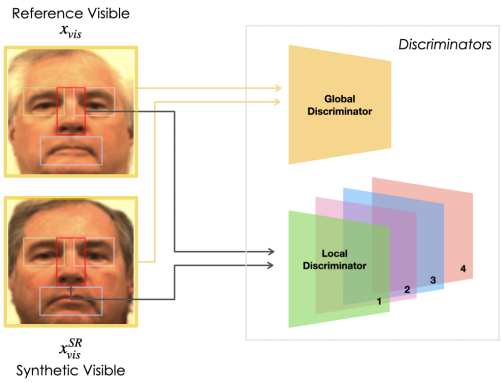

Fig. 3. **Global and Local discriminators**. While the global discriminator, applied on the whole image, is instrumental for the generator to synthesize photo-realistic HR images, the local discriminators, denoted by L_1, L_2, L_3 and L_4, focus on areas located around eyes, nose and mouth, respectively. They are designed to focus on generated details of cross-spectral biometric features.

to the HR size of 128×128. For the training phase, LR images are down-sampled from the HR thermal images with four different scale factors giving batch of images of size 128×128, 64×64, 32×32 and 16×16, respectively. Moreover, the training dataset is augmented by random sharpness, center cropping and horizontal flips.

4. EXPERIMENTAL RESULTS

4.1. Dataset and Protocol

Towards evaluating the performance of ANYRES, we train it on four benchmark multi-spectral face datasets, separately. We summarize the datasets in Table 2.

Table 2. Characteristics of datasets used for the experiments

Dataset	ARL-VTF	VIS-TH	SF	Tufts
Reference	[12]	[13]	[14]	[15]
Number of training subjects	295	40	100	50
Number of testing subjects	100	10	42	63

4.2. Evaluation and Comparison

Figure 1 and Table 1 highlight qualitative and quantitative comparison results of selected methods, additional methods can be found in supplemental material. Results are reported in terms of (i) FR biometrics standards, we present the Area Under the Curve (AUC) and Equal Error Rate (EER) metrics related to the ArcFace-based FR matcher[1]; as well as (ii) image quality evaluated by the structural similarity index measure (SSIM)[2]. We adopt the provided code for *super resolution* and related *leading* dedicated methods for comparison purpose, specifically SRGAN [16] and AxialGAN [2], respectively. While ANYRES is designed to handle *any* resolution (shortened by ANYRES-multi), we trained other methods that had been originally designed for specific (mono) resolution.

All tested methods rely on adversarial training, nevertheless they differ in the way they ensure faithful cross-spectral identity. Our first observation is that ANYRES outperforms across datasets other methods for every resolutions *w.r.t.* FR performances, and this trend is confirmed with the additional methods presented in supplemental material. Moreover, from 32×32 to 128×128, we notice that biometric performances are roughly the same across resolutions, which indicates the ability of identity-consistency through various resolutions. *W.r.t.* image quality, ANYRES depicts stable SR images across resolutions without artefacts. In almost all resolutions and datasets, it achieves either best or a close second best SSIM score as opposed to other methods. We note that CFR relies on biometric features rather than perceptual features (pixel scale), and therefore we consider image quality being secondary to biometrics performance, which we place emphasis on.

Unexpectedly, SRGAN which is originally built to super-resolve an image within the same spectrum, has demonstrated competitive results that could surely be improved, in case that specific loss functions were added. Nevertheless, its design is not optimized for accumulating resolutions and this could be explained by the residual blocks fashion processing. Finally, our approach significantly boosts the performances and demonstrates the ability to handle any resolution solely with one unified framework. Results presented on the ARL-VTF dataset significantly exceed the comparative scores. This gap is explained by the fact that, ARL-VTF is the largest thermal-visible paired face dataset publicly available, and includes over 500,000 images, unlike other datasets which contain hundred images. Note that the SF dataset contains (extreme) pose faces, which brings random variation during training. Further results confirm the ability of ANYRES to be operational in unconstrained-CFR systems.

4.3. Discussion

ANYRES benefits from *Pyramid-like* architecture, and we hereby proceed to motivate our choice. Towards addressing SR and CFR, the greatest advantage of pyramid representation is the property to convert global image features into local features, while condensing representation of the whole image. In such a process, successive levels of the pyramid become a reduced-resolution version of the image, thus relying on a multi-scale analysis. ANYRES is built solely with one single network and enables the handling of any resolution unlike AxialGAN and all other tested methods which are working from a fixed resolution.

An ablation study further confirms the positive impact of loss functions included in ANYRES. Extensive experimentation can be found in the supplemental material.

Finally, to demonstrate the versatility of the method we train ANYRES on 4 resolutions (see section 3.4). This choice stems from constraints related to the adversarial training.

[1]Higher AUC indicates better performance, whereas lower EER is better. Other evaluation metrics can be found in the supplemental material.

[2]Score of 1 is the extreme case of comparing identical images.

Table 1. Quantitative comparison on four multi-spectral face datasets. Experimental results validate accuracy *w.r.t.* facial recognition, namely by AUC % and EER % scores, as well as visual quality, as reported by SSIM %. Bold indicates the best performance.

Res.	Method	ARL-VTF dataset			VIS-TH dataset			SF dataset			Tufts dataset		
		AUC	EER	SSIM	AUC	EER	SSIM	AUC	EER	SSIM	AUC	EER	SSIM
16×16	SRGAN	82.68	25.69	50.21	69.91	36.86	49.15	71.76	34.09	56.57	50.86	49.31	14.76
	AxialGAN	84.45	22.92	59.74	72.14	33.82	53.82	77.61	29.12	64.69	59.65	43.10	38.79
	ANYRES-multi	**91.24**	**16.90**	**63.05**	**75.94**	**29.66**	**53.89**	**77.64**	**28.04**	**65.34**	**63.62**	**40.23**	**39.01**
32×32	SRGAN	95.26	12.02	56.50	84.86	22.04	54.01	82.81	24.79	64.20	58.44	44.14	23.25
	AxialGAN	95.19	11.85	**64.43**	84.95	22.50	**57.36**	84.24	23.34	67.93	63.38	40.39	39.91
	ANYRES-multi	**98.61**	**6.17**	64.23	**87.01**	**20.82**	55.35	**84.77**	**23.08**	**68.33**	**77.22**	**29.09**	**40.01**
64×64	SRGAN	97.85	5.30	59.72	88.07	20.86	56.97	86.02	21.42	63.02	66.60	38.04	33.62
	AxialGAN	97.22	9.19	66.39	88.09	20.88	**57.78**	86.14	21.49	67.33	64.23	39.69	41.31
	ANYRES-multi	**99.42**	**4.02**	**67.05**	**89.57**	**18.08**	57.43	**86.24**	**21.24**	**68.85**	**80.74**	**26.41**	**41.38**
128×128	SRGAN	98.61	5.14	59.28	88.16	20.18	57.14	87.47	20.27	67.07	67.47	37.58	39.08
	AxialGAN	97.91	9.65	66.48	88.59	20.96	57.98	87.02	20.52	67.61	66.27	38.23	40.00
	ANYRES-multi	**99.44**	**3.82**	**67.02**	**89.58**	**18.04**	**57.98**	**89.13**	**18.14**	**68.93**	**80.91**	**26.29**	**40.30**

During the first epochs, the discriminator absorbs many samples of different resolution, then overpowering the generator would have a negative effect on training. AWe here not that being adaptive to the practical CFR scenario by considering 4 scales of resolutions is enough and fully reliable for GAN training. We also have been able to confirm through further implementations that ANYRES is capable to super-resolve with great FR accuracy intermediate resolutions. However, particular attention should be paid to very LR images, where no biometric information is contained thus making impossible to unveil the visible face.

5. CONCLUSIONS

CFR systems necessitate accurate and reliable automated models, able to handle a wide range of resolutions. In this paper, we proposed ANYRES, a unified generative model that accepts facial images of a wide range of resolutions as input, and that proceeds to accurately translate such from one spectrum to another, while ensuring faithful cross-spectral identity. Experiments on four datasets suggest that our proposed ANYRES outperforms state-of-the-art methods, even under pose variation. While this is a first step to rendering CFR systems adaptive to real world scenarios, future work will involve more steps in this direction, keeping in mind the goal of a reliable monitoring systems in unconstrained environments.

6. REFERENCES

[1] D. Anghelone, C. Chen, A. Ross, and A. Dantcheva, "Beyond the visible: A survey on cross-spectral face recognition," *preprint*, 2022.

[2] R. Immidisetti, S. Hu, and V. M. Patel, "Simultaneous face hallucination and translation for thermal to visible face verification using axial-gan," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*.

[3] C. Chen, D. Anghelone, P. Faure, and A. Dantcheva, "Attention-guided generative adversarial network for explainable thermal to visible face recognition," in *2022 IEEE International Joint Conference on Biometrics (IJCB 2022)*.

[4] D. Anghelone, C. Chen, P. Faure, A. Ross, and A. Dantcheva, "Explainable thermal to visible face recognition using latent-guided generative adversarial network," in *2021 IEEE International Conference on Automatic Face and Gesture Recognition*.

[5] 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 *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition*.

[6] T. Zhang, A. Wiliem, S. Yang, and B. Lovell, "Tv-gan: Generative adversarial network based thermal to visible face recognition," in *2018 international conference on biometrics (ICB)*.

[7] C. Chen and A. Ross, "Matching thermal to visible face images using a semantic-guided generative adversarial network," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on CVPR*, 2019.

[10] S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *2021 IEEE International Conference on Engineering and Emerging Technologies*.

[11] D. Anghelone, S. Lannes, V. Strizhkova, P. Faure, C. Chen, and A. Dantcheva, "Tfld: Thermal face and landmark detection for unconstrained cross-spectral face recognition," in *2022 IEEE International Joint Conference on Biometrics (IJCB 2022)*.

[12] D. Poster et al., "A large-scale, time-synchronized visible and thermal face dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[13] K Mallat and J-L Dugelay, "A benchmark database of visible and thermal paired face images across multiple variations," in *International Conference of the Biometrics Special Interest Group, BIOSIG 2018, Darmstadt, Germany, September*.

[14] M. Abdrakhmanova et al., "Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams," *Sensors*, 2021.

[15] K. Panetta et al., "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[16] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE conference on computer vision and pattern recognition*, 2017.