Intraclass Data Augmentation Framework for Data-Imbalanced Cross-Domain Face Recognition

Patricio Xavier Moreno-Vallejo
Escuela Superior Politécnica del Litoral (ESPOL)
Guayaquil, Ecuador
Escuela Superior Politécnica de Chimborazo (ESPOCH)
Riobamba, Ecuador
Email: pxmoreno@espol.edu.ec
ORCID: 0000-0002-9317-9884

Gisel Katerine Bastidas-Guacho
Escuela Superior Politécnica de Chimborazo (ESPOCH)
Riobamba, Ecuador
Email: gisel.bastidas@espoch.edu.ec
ORCID: 0000-0002-6070-7193

Angel D. Sappa
Computer Vision Center (CVC)
Barcelona, Spain
Escuela Superior Politécnica del Litoral (ESPOL) Guayaquil, Ecuador
Email: asappa@cvc.uab.es
ORCID: 0000-0003-2468-0031

2025 IEEE Ninth Ecuador Technical Chapters Meeting (ETCM) | 979-8-3315-5264-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ETCM67548.2025.11304378

***Abstract*—This work introduces an intraclass data augmentation framework that leverages Pivotal Tuning Inversion to augment face data by modifying attributes such as pose, expression, age, and hairstyle. The framework also incorporates a cross-domain image translation network to generate thermal infrared images from visible ones, enabling the creation of paired registered visible-infrared augmented datasets. Experimental results show that the proposed framework enhances the quality and fidelity of the generated images, as reflected by metrics such as FID, KID, LPIPS, SSIM, and PSNR. Additionally, the use of the proposed framework to generate visible-thermal infrared augmented datasets enhances the performance of existing computer vision applications, such as heterogeneous face recognition.**

***Keywords*—Data Augmentation, Cross-Domain, Face recognition, Deep Learning**

I. Introduction

The electromagnetic spectrum spans from very long radio waves to very short gamma rays, but humans can only perceive a small portion of it - visible light. While visible light can convey significant information, it is sensitive to lighting conditions. As a result, systems that rely on visible images can become less effective or even obsolete in low-light situations. This challenge can be addressed through the use of the infrared (IR) spectrum domain, which spans wavelengths from 8 to 15 microns (μm). The IR spectrum domain can be categorized into near-infrared (NIR), mid-infrared, and far-infrared. Sensors that capture IR radiation can convert this energy into images, which are then interpreted by computer vision algorithms. In this context, IR imagery can enhance existing computer vision applications, such as face recognition (FR) [1], [2], making them more robust in low-light conditions, where traditional visible-light sensors may fail to capture relevant scene details.

IR imaging is also widely used in fields such as surveillance, where it enables scene monitoring in complete darkness by detecting infrared radiation. Additionally, IR imaging in biometric systems helps to prevent face spoofing attacks by identifying subjects behind masks used by intruders or differentiating between photographs and real people's identity impersonation

Fig. 1. The proposed intraclass data augmentation framework

because portraits do not emit IR radiation. Furthermore, IR sensors improve pedestrian detection in autonomous vehicles and enhance safety in low-light environments. For instance, in [3], Dasgupta et al. exploit the information of RGB and thermal images throughout two Recurrent Neural Networks (RNNs) that process the images in order to get the bounding boxes of each pedestrian. Moreover, in agriculture, infrared images are used to do a multimodal analysis for detecting vegetation stress [4].

On the other hand, face recognition systems based on deep learning approaches have become more ubiquitous because they are found in social networks, automatic biometric identification, national security, and forensic systems. This broad use has been possible given that currently, there are large databases of face images in the visible spectrum, which have allowed exploiting deep learning solutions to create quite robust systems in recognizing faces in this spectrum. While there are datasets with millions of data in the visible domain, other domains only have hundreds or thousands of data. This limitation makes it difficult for deep learning approaches to extract and capture the features of images in different domains to determine their distribution. Despite that, in recent years, the price of IR cameras has considerably dropped so much that they can already be found in portable devices such as laptops or cell phones. Because of this, some databases already include images of faces in the visible spectrum and their corresponding images in the infrared spectrum. These databases have opened the door to expanding the horizons of face recognition systems beyond the visible.

An approach to extend existing visible face recognition systems to work with NIR or IR imagery is synthesizing non-visible images to the visible spectrum. Although IR can

979-8-3315-5264-0/25/$31.00 ©2025 IEEE

capture images in the dark, some details in the face are not captured by these sensors, making it challenging to synthesize the visible face as close to the real image. Another challenge is the varying poses because some approaches only work well when all images are frontal faces. However, when different poses are added, they have a low performance. Fondje et al. [5] propose a domain and pose invariant framework that uses networks to extract the features of visible and IR and transform them to bridge the domain and pose gaps. Some approaches, like heterogeneous face recognition, try to bridge the gap between different domains in a latent space to allow the system to work with any domain. Deep learning approaches have been used to achieve this goal because they can capture non-linear relationships, allowing them to identify complex patterns [6]. However, the lack of registered data in different domains can make it difficult for the models to capture the differences and equalities between domains, and there are models like diffusion models that require registered data.

A way to overcome the limitation of registered images in multiple domains is by using data augmentation [7]. Most of the best techniques to perform data augmentation include deep learning approaches that, in most cases, need at least 8000 samples during training to perform well. This amount of images is challenging to get and more difficult to register with the visible images. Another approach is to use loss functions to avoid the overfitting of the model, allowing the model to work with few data for training [8].

Two main approaches are commonly used for face data augmentation: intraclass and interclass data augmentation. Intraclass data augmentation works on modifying existing images to increase variety while keeping the original identities intact. The changes applied in intraclass data augmentation include rotation, scaling, adjusting colors, or adding synthetic noise, allowing the images to represent the same identity under different conditions. In contrast, interclass data augmentation broadens the dataset by generating entirely new synthetic identities that share characteristics with the original dataset. Techniques like generative adversarial networks (GANs) or diffusion models are often used in both approaches to create synthetic images, ensuring they follow the same distribution as the original data but represent completely new identities (interclass) or keep the same input identity (intraclass).

The current work investigates the challenge of cross-domain data augmentation within the context of intraclass variations, addressing the issue of limited data across domains. It uses deep-learning techniques with relatively few samples. To this end, the proposed network is designed and trained to take visible spectrum images as input and produce corresponding thermal infrared images as output by using CycleGAN [9] with a ResNet backbone. Thermal infrared refers specifically to long-wavelength infrared radiation. Additionally, Pivotal Tuning Inversion (PTI) [10] is used to perform data augmentation on visible images based on the adaptation of the latent space of pre-trained robust models, such as StyleGAN or StyleCLIP. The output from PTI is then fed into the proposed trained network in order to generate the corresponding infrared images in the augmented dataset, as shown in Fig. 1. The primary goal of this approach is to bridge the gap between domains with limited data by augmenting datasets in a way that preserves the critical characteristics needed for downstream tasks.

This document is organized as follows: Section II provides an overview of related work. Section III outlines the proposed method, detailing the framework and techniques employed. Section IV describes the dataset and presents the analysis of the experimental results. Finally, Section V summarizes the findings and concludes the work.

II. Related work

When referring to the use of cross-domain data in computer vision (CV) tasks, the term 'heterogeneous' is often added to the names of well-known CV tasks to emphasize the disparity between the data. For example, the task of heterogeneous face recognition implies that the images in the gallery and the probe belong to different domains. Data augmentation techniques have been proposed to bridge the gap between domains. For instance, Ye et al. [11] proposed an intraclass channel-augmented joint learning framework to address the challenges of visible-infrared recognition, particularly in scenarios involving cross-modality matching, such as visible-infrared person re-identification (VI-ReID) and heterogeneous face recognition. Their approach includes randomly swapping RGB color channels to improve resilience to color variations, maintaining the original network structure unchanged. Beyond channel augmentation, the authors introduced a channel-level random erasing technique to simulate occlusions and improve generalizability. They also developed a channel-augmented joint learning strategy, treating channel-augmented images as an auxiliary modality to explicitly optimize cross-modality learning.

Josi et al. [12] addressed the challenge of V-I ReID under real-world conditions, where image corruption (e.g., noise, blur, and weather effects) significantly impacts model performance. They proposed an intraclass multimodal data augmentation (MDA) strategy to leverage the complementary information of visible (RGB) and infrared (IR) modalities. Their approach introduces techniques like Multimodal Soft Random Erasing (MS-REA), which applies localized random erasing to both modalities and Multimodal Patch Mixing (M-PATCH), where patches from one modality are superimposed on the other to encourage cross-modal feature learning. However, applying this MDA strategy to face datasets can be challenging due to the higher sensitivity of facial recognition tasks to spatial distortions and occlusions. Unlike person re-identification datasets, where the human body offers multiple regions for feature extraction, face datasets rely heavily on preserving fine-grained, modality-specific details (e.g., eye or mouth regions). Techniques like MS-REA and M-PATCH may inadvertently degrade critical facial features, reducing the effectiveness of cross-modal learning and introducing biases in feature representation.

In [13], Yu et al. propose CMOS-GAN, a semi-supervised cross-modality face synthesis model that consists of a face

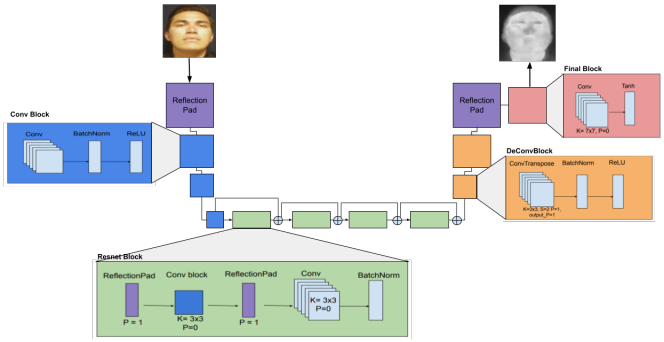

Fig. 2. The proposed CycleGAN-based network architecture for generating cross-domain image translation.

image synthesis network and a discriminator network. This model allows the synthesis of VIS images from NIR images, as well as other cross-modality syntheses such as RGB-to-depth and sketch-to-photo. Another approach to reducing the gap in cross-domain images is the featured-based domain adaptation, which means the mapping of the data from different distributions to a pre-defined space where the distributions are similar, preserving the data properties. In the last decade, the approaches have moved to deep learning architectures, specifically on variations of the Generative Adversarial Networks (GAN) and diffusion models.

The vast diversity of domains has motivated the development of coupled neural networks that work together to bridge the gap between domains. In this context, [14] is the first time that the coupling of hidden units in an AutoEncoder (AE) network was used for heterogeneous face recognition. In order to use this, one needs to assume that there exists a nonlinear common subspace where the gap between domains is bridged in such a way that similar identities have similar representations, regardless of the domain. In addition to the stacked coupling constraint, the authors use the reconstruction constraint to guide the training phase and the regularization function Kullback-Leibler Divergence (KLD) to prevent overfitting. A similar approach is applied in [15] with the difference that they use a GAN architecture instead of coupling two AE networks.

Since the appearance of GANs, most approaches have used this architecture to work with cross-domain images. Thus, it is imperative to delve into this topic by reviewing the improvements made in recent years. First, Lee et al. [16] propose a GAN-based cross-domain unpaired image-to-image translation with one encoder for each domain and a single decoder. The encoder for the RGB images encapsulates the edges, while the Thermal Infrared (TIR) encoder encapsulates the pixel intensity, enhancing the structural consistency in cross-domain transformations. Beyond the adversarial and cycle losses used in GANs, they use a Laplacian of Gaussian (LoG) loss to enforce texture and edge consistency. This approach improves domain adaptation by maintaining semantic coherence and minimizing distortions in synthesized images.

Güzel and Yavuz [17] explore thermal image synthesis from RGB images using CycleGAN, emphasizing the importance of paired image training and electromagnetic spectrum-based normalization. Expanding on this, Dubey et al. [18] propose a conditional GAN (cGAN) model for TIR colorization, adapting the pix2pixHD GAN architecture with an Earth Mover's Distance and total variation loss to enhance color fidelity and reduce artifacts. Furthermore, Sigillo et al. [19] propose a structural-aware generative adversarial network (StawGAN) designed for infrared-to-RGB image translation, focusing on improving the quality and structure of generated images rather than simple colorization. Unlike standard GAN-based approaches, StawGAN incorporates a shared encoder-decoder architecture that simultaneously processes the image and its segmented target, enforcing structural consistency. Additionally, it incorporates a contrast network that assists the generator in selecting appropriate contrast, sharpness, and gamma levels.

Besides, Wang et al. [20] proposes CannyGAN, which embeds the classic canny edge detection to a CycleGAN to synthesize VIS images from IR images. This approach aims to increase the texture of input IR images in a latent space embedding with the Canny edges. This strategy helps the network's training to achieve better results than the simple CycleGAN.

Recent research has explored generative models for data augmentation. One such approach is the Edge-Guided Conditional Diffusion Model (ECDM), which synthesizes high-quality, pixel-aligned pseudo-thermal images from visible images by leveraging edge information as contextual guidance [21]. ECDM employs a two-stage adversarial training strategy to eliminate domain-specific inconsistencies while preserving essential structural details. Additionally, Mayr et al. [22] introduce TIR ControlNet, a diffusion-based model specifically re-trained for thermal infrared image synthesis. Their method conditions the image generation process on existing semantic segmentation maps, allowing the creation of highly realistic synthetic TIR images without requiring manually labeled datasets. These advances in conditional image synthesis improve thermal image dataset diversity and enhance deep learning performance in low-data scenarios.

In general, deep learning models require a large amount of data to learn and generalize. Nevertheless, the data in domains beyond the visible are limited. Hence, a common approach to overcome this limitation is data augmentation. The literature includes works claiming that their approaches perform effectively with few-shots like [8] where Ojha et al. train a GAN network using as few as ten samples of the target domain. This approach uses a cross-domain distance consistency loss that helps to preserve the similarities and differences in images in the source and an anchor-based strategy on the discriminator to reduce overfitting. Otherwise, Fu et al. [23] uses large VIS datasets to enrich the diversity of the images to be used in heterogeneous face recognition. This approach is made up of three parts. First, they train the generators for each domain with limited paired data to learn to preserve the identity of the generated pairwise data. Second,

they train the generators with unpaired data but inject VIS data from a large dataset to introduce diversity in the identities generated. Third, the generators create a large synthetic dataset of paired heterogeneous data to train the Heterogeneous Face Recognition (HFR) network.

III. Proposed method

A. Problem definition

Cross-domain face recognition in imbalanced datasets presents challenges due to the lack of paired visible-infrared data and the domain shift between these modalities. This work aims to generate augmented face datasets by modifying intra-class variations in the visible spectrum and translating these variations into the infrared domain while preserving identity and structural consistency. Formally, given a visible image I_{vi}, the objective is to synthesize a corresponding thermal infrared image I_{tir}, ensuring consistency across both domains:

$$I_{\text{tir}} = G(A(I_{\text{vi}})), \tag{1}$$

where $A(\cdot)$ consists of an intraclass data augmentation network that generates diverse variations of I_{vi} before translation. $G(\cdot)$ represents the cross-domain image translation network, which must learn to retain identity features while transferring the spectral characteristics of the thermal domain, preserving details such as facial features, pose, expression, and hairstyle.

B. Intraclass Data Augmentation Framework

The proposed framework is referred to as the intraclass data augmentation framework. It consists of a cross-domain image translation network and a visible-data augmentation network. The cross-domain image translation network is based on CycleGAN, utilizing a ResNet backbone. The network takes images of faces in the visible domain as input and transforms them into the thermal infrared domain while preserving details such as position, expression, facial features, and hairstyle as shown in Fig. 2. While CycleGAN typically requires 10k images for training, the proposed approach achieves this with only 1.5k pairs of images by incorporating a distance consistency loss [8].

The distance consistency loss ensures that the distribution of the cosine similarity between selected feature maps of the generator transforming images from the visible domain to the thermal infrared domain (v-tir) is similar to that of the generator transforming images from the thermal infrared domain to the visible domain (tir-v). This mechanism helps prevent overfitting by forcing the networks to preserve identity similarities and differences across both domains, avoiding the repetitive generation of identical subjects. This loss is defined as follows:

$$\mathcal{L}_{\text{dist}}(G_{v\to tir}, G_{tir\to v}) = \mathbb{E}_{\{z_i\sim p_z(z)\}} \sum_{l,i} D_{\text{KL}}\left(\underset{l_c}{\text{sim}}(y_i^{v\to tir,l}, y_i^{vis\to tir,l_c}) \| \underset{l_c}{\text{sim}}(y_i^{tir\to v,l}, y_i^{tir\to v,l_c}) \right), \tag{2}$$

where $G_{v\to tir}$ and $G_{tir\to v}$ are the generator models that produce the feature maps $y_i^{v\to tir}$ and $y_i^{tir\to v}$, respectively. The index l represents the feature map produced by layer l of the respective generator network. On the other hand, l_c is used to extract the feature map produced by the same layer l, but for the other elements in the batch, enabling a comparison of feature representations between different samples in the batch. sim_{l_c} calculates the cosine similarity between the feature maps of the anchor layer l and their respective feature maps on layers l_c from the other elements of the batch. This similarity measures how closely the generated feature representations align with each other. Additionally, D_{KL} measures the Kullback-Leibler divergence, quantifying the difference between the probability distributions of feature similarities computed after applying the softmax function.

To get the similarities of each model, a batch of n real visible images is selected, with n set to 4 in this study. This batch serves as input to the v-tir model, which transforms visible images into infrared. The output consists of a batch of generated thermal infrared images and the intermediate feature maps of the v-tir model. The generated infrared images are then passed to the tir-v model, which transforms them back to the visible domain. However, in this case, only the intermediate feature maps from the tir-v generator are retained for further calculations. Once intermediate feature maps from both models (v-tir and tir-v) are obtained, a random anchor feature map is selected and compared with its corresponding one in the rest of the batch to get the cosine similarity. This process is then repeated by selecting a different random anchor feature map on each iteration until at least one feature map of each input in the batch is compared.

After training the cross-domain image translation network, data augmentation of the real images in the visible domain is performed within the proposed framework using the technique of Pivotal Tuning Inversion (PTI) [10]. PTI allows the modification of the generator of a StyleGAN model pre-trained with the Flickr-Faces-HQ (FFHQ) dataset to keep the identity of the original input image without affecting the model's high editing capabilities. Once the generator of the StyleGAN has been fine-tuned, it is used to generate new images of the subjects in the database with different poses, keeping the identity. The same procedure is performed with the StyleCLIP model pre-trained on FFHQ to generate images with variations in hairstyle, age, and expression. Then, the augmented images in the visible domain are subsequently transferred to the thermal domain using the proposed cross-domain image translation network. This process enables the generation of registered data in both domains.

IV. Experiments

In this section, the dataset and experimental results are detailed to show the effectiveness of the proposed framework.

A. Dataset

The experiments in this work use the Tufts Face Database [24], focusing specifically on images from visible and thermal infrared domains. The dataset includes a total of 1,531 registered images from 112 identities. For training, 79 identities are used, comprising 1,079 images, while 11 identities with 153 images are reserved for validation, and 22 identities with 299 images are allocated for testing. This distribution ensures a balanced evaluation across training, validation, and testing phases.

TABLE I
Quantitative comparison of the original CycleGAN and CycleGAN with distance consistency loss

Experiment	FID Score ↓	KID Score ↓	LPIPS ↓	SSIM ↑	PSNR ↑
Original CycleGAN	130.2228	0.1183	0.2841	0.7833	15.6435
Ours	**85.4463**	**0.0548**	**0.2437**	**0.8101**	**16.2256**

TABLE II
TEST ACCURACY OF THE HETEROGENEOUS FACE RECOGNITION OFF-THE-SHELF APPROACH TRAINED WITH VARYING LEVELS OF DATA AUGMENTATION AND TESTED WITH 277 IMAGES. HIGHER AUGMENTATION LEVELS IMPROVE THE TOP RECOGNITION PERFORMANCE.

Augmentation (VIS-IR pairs)	Total Training Image VIS-TIR pairs	Top-1 (%)	Top-3 (%)	Top-5 (%)
Baseline:				
No data augmentation (real 1 pair/identity)	22	28.88	51.26	64.22
Interclass augmentation = **264** [23]	286	43.32	61.73	72.20
Intraclass data augmentation:				
3 pairs per identity = 66	88	29.96	50.54	62.09
6 pairs per identity = 132	154	37.55	63.18	72.20
9 pairs per identity = 198	220	39.71	59.93	72.20
12 pairs per identity = **264**	286	**46.93**	**64.98**	**73.65**

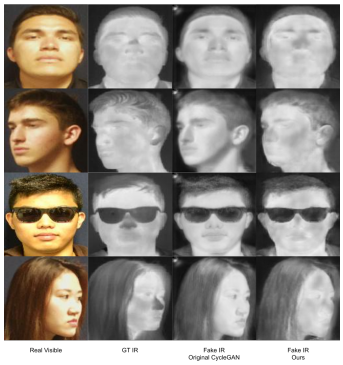

Fig. 3. Comparison of the infrared image generation results of the original CycleGAN model (third column) with those of the proposed model (right column). For reference, the two left columns display the original pair of images.

B. Results

In order to evaluate the proposed framework, we firstly assess the quality of the infrared images produced by the proposed cross-domain image translation network using as metrics Fréchet inception distance (FID) [25], Kernel Inception Distance (KID) [26] Learned Perceptual Image Patch Similarity (LPIPS) [27], Peak signal-to-noise ratio (PSNR), and Structural similarity (SSIM). The results are quantitative compared with the results of the original CycleGAN network. To this end, the original CycleGAN network is also trained using the Tufts Face Database [24] and is tested with 299 images from 22 identities. Additionally, images are generated using the method proposed in [6]; however, since this method creates new identities instead of transforming existing subjects, no ground truth is available to measure the quality of the generated images. As shown in Table I, the proposed cross-domain image translation network achieves better results. The FID and KID scores decrease, which indicates better alignment between the generated and real thermal infrared image distributions. Similarly, perceptual similarity improves with a reduction in LPIPS. At the same time, SSIM increases, showing structural fidelity. Additionally, the PSNR improves, reflecting reduced reconstruction error. These results highlight the effectiveness of the proposed cross-domain image translation network across all evaluated metrics and indicate that adding the distance consistency loss enhances the quality, structural preservation, and perceptual accuracy of the infrared-generated images.

The qualitative results in Fig. 3 illustrate a visual comparison between the original CycleGAN model and the proposed model with distance consistency loss. The infrared images generated by the original CycleGAN exhibit artifacts and resemble simple grayscale images, losing the distinct characteristics of infrared images. In contrast, the proposed model produces infrared images with improved visual quality, accurately preserving structural features and finer details. The proposed model shows better alignment with the ground-truth thermal infrared images, particularly in challenging cases such as images with sunglasses or side profiles.

To demonstrate the effectiveness of the data augmentation approach for improving existing computer vision applications, such as FR, an off-the-shelf heterogeneous FR method from [6] is used. This method treats FR as a classification task and employs the Light CNN framework [28] as its backbone. The gallery is constructed using one frontal face image per subject in each domain from the test dataset, which consists of 22 pairs of registered images. The probe set includes the remaining thermal infrared images in the test dataset, totaling 277 images, and is used to evaluate the performance of the approach. The heterogeneous FR model is first trained using the gallery with images in both domains and then tested with the probe images in the infrared domain. The procedure is repeated with the addition of augmented data. The data augmentation process follows the details indicated in Table II. First, no augmentation is applied, and the training set consists only of the original 22 pairs of registered images, achieving a baseline top-1 accuracy of 28.88%. Additionally, the interclass data augmentation method proposed in [23] was used as a baseline, testing it with the current dataset protocol, achieving a top-1 accuracy of 43.32% with the augmentation of 264 new pairs of identities. To make a fair comparison, the number of augmented pairs matches the highest-performing intraclass data augmentation setting. Then, the proposed framework was used to augment data to the original 22 images adding progressively 66 additional pairs (3 pairs per identity), 132 additional pairs (6 pairs per identity), 198 additional pairs (9 pairs per identity), and finally, 264 additional pairs (12 pairs per identity). The corresponding total training datasets consist of 88, 154, 220, and 286 VIS-TIR pairs, respectively. As the amount of augmented data increases, the test accuracy improves, reaching 46.93% when 12 pairs per identity are included in the training set. This demonstrates the significant impact of data augmentation in improving the heterogeneous face recognition performance. Additionally, the results show that when augmenting data using intraclass data augmentation, the FR recognition model performs better, achieving a higher top-1, top-3, and top-3 accuracy than when augmenting the same amount of data using an interclass data augmentation model. This experiment also provides evidence that the proposed framework for data augmentation preserves critical identity features across both visible and infrared domains since the accuracy of the FR model improves when adding augmented data.

Fig. 4 shows how the proposed framework uses Pivotal Tuning Inversion (PTI) for intraclass data augmentation in order to generate variations in pose, expression, and hairstyle while preserving identity. The corresponding thermal images are synthesized using the CycleGAN-based cross-domain image translation network. In that way, the proposed approach enhances the diversity of the dataset while maintaining feature integrity to improve the performance of heterogeneous face recognition models.

V. CONCLUSIONS

The proposed intraclass data augmentation framework incorporates PTI to augment data by modifying attributes such as pose, expression, age, and hairstyle. Additionally, our framework integrates the cross-domain image translation network, which effectively generates infrared images from visible ones, enabling the creation of paired visible and infrared datasets. This network uses the distance consistency loss to keep the distributions of similarities from TIR-VIS and VIS-TIR models as close as possible, ensuring that the important details from the visible domain are transferred to the infrared domain and avoiding overfitting. Moreover, the proposed framework demonstrates its effectiveness in improving the performance of existing computer

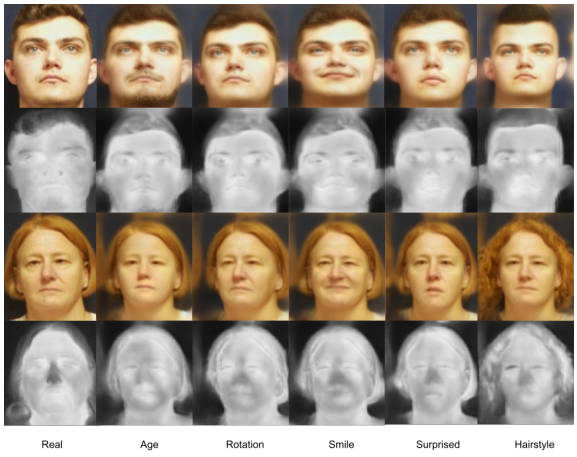

Fig. 4. Examples of augmented face images generated using the proposed intraclass data augmentation framework. The first column contains the original images, followed by augmented variations modifying attributes such as pose, expression, and hairstyle

vision applications, such as heterogeneous face recognition, since it provides augmented and paired datasets, which enhance model training and generalization. For future work, the framework could be extended by exploring the integration of advanced generative models, such as diffusion models, as the backbone for the cross-domain image translation network. Diffusion models have demonstrated superior performance in generating high-quality and diverse samples, which could further improve the quality of generated infrared images and paired datasets.

REFERENCES

[1] 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)*. IEEE, 2022, pp. 1–9.

[2] Y. Mei, P. Guo, and V. M. Patel, "Escaping data scarcity for high-resolution heterogeneous face hallucination," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 676–18 686.

[3] K. Dasgupta, A. Das, S. Das, U. Bhattacharya, and S. Yogamani, "Spatio-Contextual Deep Network-Based Multimodal Pedestrian Detection for Autonomous Driving," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[4] R. Adhitama Putra Hernanda, H. Lee, J. il Cho, G. Kim, B.-K. Cho, and M. S. Kim, "Current trends in the use of thermal imagery in assessing plant stresses: A review," *Computers and Electronics in Agriculture*, vol. 224, p. 109227, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169924006185

[5] C. N. Fondje, S. Hu, and B. S. Riggan, "Learning domain and pose invariance for thermal-to-visible face recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 1, pp. 15–28, 2022.

[6] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "Dvg-face: Dual variational generation for heterogeneous face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2938–2952, 2021.

[7] M. M. Moradi and R. Ghaderi, "I-gans for synthetical infrared images generation," in *2022 International Conference on Machine Vision and Image Processing (MVIP)*. IEEE, 2022, pp. 1–6.

[8] U. Ojha, Y. Li, J. Lu, A. A. Efros, Y. J. Lee, E. Shechtman, and R. Zhang, "Few-shot Image Generation via Cross-domain Correspondence," *CVPR*, pp. 10 743–10 752, 2021. [Online]. Available: http://arxiv.org/abs/2104.06820

[9] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 2242–2251, 12 2017.

[10] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal Tuning for Latent-based Editing of Real Images," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 1, pp. 1–13, 8 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3544777

[11] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13 547–13 556.

[12] A. Josi, M. Alehdaghi, R. M. O. Cruz, and E. Granger, "Multimodal data augmentation for visual-infrared person reid with corrupted data," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2023, pp. 1–10.

[13] S. Yu, H. Han, S. Shan, and X. Chen, "Cmos-gan: Semi-supervised generative adversarial model for cross-modality face image synthesis," *IEEE Transactions on Image Processing*, vol. 32, pp. 144–158, 2023.

[14] B. S. Riggan, C. Reale, and S. Member, "Coupled Auto-Associative Neural Networks for Heterogeneous Face Recognition," *IEEE Access*, vol. 3, pp. 1620–1632, 2015.

[15] S. M. Iranmanesh, B. Riggan, S. Hu, and N. M. Nasrabadi, "Coupled generative adversarial network for heterogeneous face recognition," *Image and Vision Computing*, vol. 94, p. 103861, 2020. [Online]. Available: https://doi.org/10.1016/j.imavis.2019.103861

[16] D. Lee, M. Jeon, Y. Cho, and A. Kim, "Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8291–8298.

[17] S. Güzel and S. Yavuz, "Infrared image generation from rgb images using cyclegan," in *2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 2022, pp. 1–6.

[18] E. Dubey, N. Singh, P. Joshi, and R. Prasad, "A conditional gan architecture for colorization of thermal infrared images," in *2023 IEEE World AI IoT Congress (AIIoT)*, 2023, pp. 0055–0062.

[19] L. Sigillo, E. Grassucci, and D. Comminiello, "Stawgan: Structural-aware generative adversarial networks for infrared image translation," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2023, pp. 1–5.

[20] T. Wang, T. Zhang, L. Liu, A. Wiliem, and B. Lovell, "CannyGAN: Edge-Preserving Image Translation with Disentangled Features," *Proceedings - International Conference on Image Processing, ICIP*, vol. 2019-September, pp. 514–518, 9 2019.

[21] G. Zhu, H. Pan, Q. Wang, C. Tian, C. Yang, and Z. He, "Data generation scheme for thermal modality with edge-guided adversarial conditional diffusion model," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 10544–10553. [Online]. Available: https://doi.org/10.1145/3664647.3680922

[22] C. Mayr, C. Kubler, N. Haala, and M. Teutsch, "Narrowing the synthetic-to-real gap for thermal infrared semantic image segmentation using diffusion-based conditional image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 3131–3141.

[23] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He, "DVG-Face: Dual Variational Generation for Heterogeneous Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. XX, no. X, 2021.

[24] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 509–520, 2018.

[25] G. Parmar, R. Zhang, and J.-Y. Zhu, "On aliased resizing and surprising subtleties in gan evaluation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 410–11 420.

[26] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[28] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *IEEE transactions on information forensics and security*, vol. 13, no. 11, pp. 2884–2896, 2018.