Synthesis-Guided Feature Learning for Cross-Spectral Periocular Recognition

Domenick Poster and Nasser Nasrabadi
Lane Dept. of Computer Science and Electrical Engineering, West Virginia University, Morgantown, USA

Abstract— **A common yet challenging scenario in periocular biometrics is cross-spectral matching - in particular, the matching of visible wavelength against near-infrared (NIR) periocular images. We propose a novel approach to cross-spectral periocular verification that primarily focuses on learning a mapping from visible and NIR periocular images to a shared latent representational subspace, and supports this effort by simultaneously learning intra-spectral image reconstruction. We show the auxiliary image reconstruction task (and in particular the reconstruction of high-level, semantic features) results in learning a more discriminative, domain-invariant subspace compared to the baseline while incurring no additional computational or memory costs at test-time. The proposed Coupled Conditional Generative Adversarial Network (CoGAN) architecture uses paired generator networks (one operating on visible images and the other on NIR) composed of U-Nets with ResNet-18 encoders trained for feature learning via contrastive loss and for intra-spectral image reconstruction with adversarial, pixel-based, and perceptual reconstruction losses. Moreover, the proposed CoGAN model beats the current state-of-art (SotA) in cross-spectral periocular recognition. On the Hong Kong PolyU benchmark dataset, we achieve 98.65% AUC and 5.14% EER compared to the SotA EER of 8.02%. On the Cross-Eyed dataset, we achieve 99.31% AUC and 3.99% EER versus SotA EER of 4.39%.**

I. INTRODUCTION

Periocular recognition uses the region surrounding the eye to recognize individuals. The periocular region can offer a good trade-off in terms of recognition accuracy versus useability. Cross-spectral recognition further relaxes the constraints imposed on input images at the cost of introducing a challenging domain gap between images of different spectral wavelengths. However, this proposition is made more attractive for periocular recognition due to the prevalence of iris recognition systems which typically generate near-infrared (NIR) periocular images, and also to the increased use of masks being worn which partially obscure the face. Examples of periocular images are shown in Fig. 1.

Two common strategies to address the cross-spectral domain-shift are to 1) map images of either domain to a shared representational subspace [1], [2], [7], [18], or 2) translate images from one domain to the other [8], typically to use an existing uni-modal recognition model. Drawing on elements of both strategies, we propose a novel approach to cross-spectral periocular recognition which utilizes coupled convolutional neural networks (CNN) to perform shared latent subspace learning and feature matching while simultaneously leveraging Conditional Generative Adversarial Networks [4], [12] (cGAN) to guide the learning process. Our Coupled GAN (CoGAN) architecture outperforms the state-of-the-art on the Hong Kong PolyU [13] and Cross-Eyed [15] cross-spectral periocular recognition datasets.

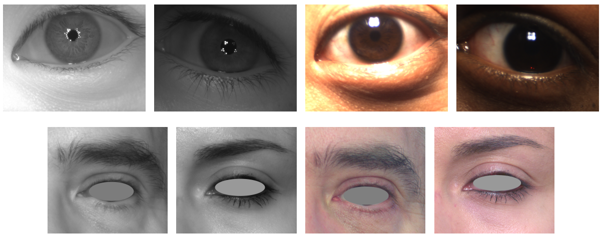
Fig. 1: Near-infrared (left) and visible (right) periocular images from HK PolyU (top) and Cross-eyed (bottom) datasets.

The primary contributions of this work are:

- A novel cross-spectral periocular recognition approach utilizing the proposed CoGAN architecture.
- A set of benchmark experiments on the Hong Kong PolyU and Cross-Eyed cross-spectral periocular datasets validating the performance of the proposed approach over baseline frameworks and state-of-the-art.
- A series of ablation studies validating the benefits of the auxiliary goal of image reconstruction to shared subspace feature learning.

II. RELATED WORKS

Deep-learning based approaches have been used in a variety of periocular biometric scenarios including intra-spectral recognition [6], [7], [11], [18], bi-modal fusion [18], and cross-spectral recognition [1], [8], [18]. In cross-spectral periocular recognition, near perfect or perfect results have been attained [1], [8], [18] on the Hong Kong PolyU [13] and Cross-Eyed [15] *closed world* protocols due to the high correlation between the *non-class-disjoint* train and test data. However, Zanlorensi *et al.* [18] also achieve state-of-the-art performance on the more challenging *class-disjoint open world* protocol using coupled ResNet-50 [5] networks trained for feature extraction and matching. Our approach primarily differs in that we enhance the subspace feature learning by simultaneously training for image reconstruction in a multi-task fashion while also using a much smaller ResNet-18 [5] network.

A generative adversarial network (GAN) [4] is composed of a generator network $G(z)$ which generates data given a random input vector z, and a discriminator network $D(\cdot)$ which outputs the probability that the input came from the

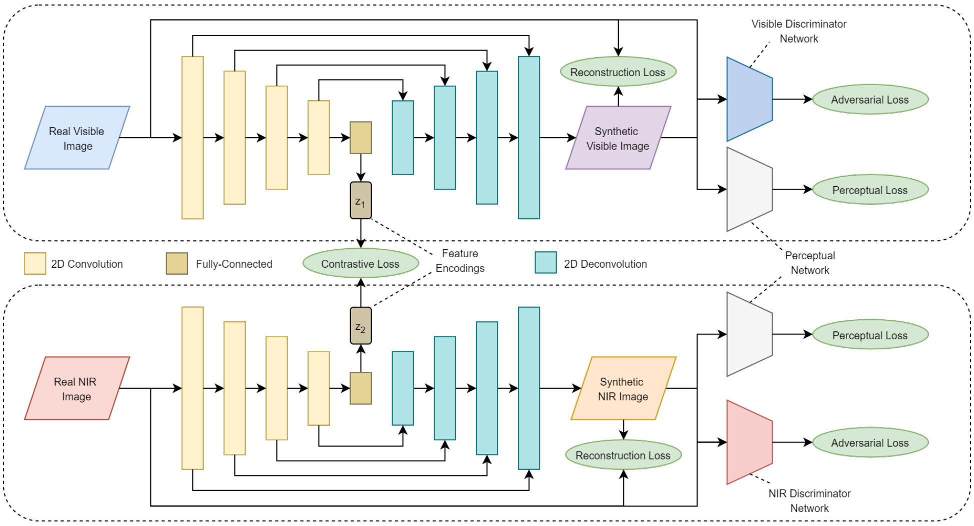

Fig. 2: The CoGAN architecture composed of coupled ResNet-18 encoders (yellow) embedded within dual U-Net cGANs.

training data x or from the generator. These two networks compete by playing a "minimax game." In a cGAN [12], the networks are conditioned on some auxiliary information y. The generator function takes the form $G(z|y)$ instead of $G(z)$. The adversarial loss for the cGAN is:

$$\min_G \max_D \mathcal{L}_{cGAN}(G;D;x;y) = E_{x\sim P_{data}(x)}[log(D(x|y))] + E_{z\sim P_z(z)}[log(1-D(G(z|y)))]. \quad (1)$$

Several methods have been proposed to refine the quality of the generated images. U-Nets [14] are generators with skip connections that forward lower-level features from the encoder layers to their mirrored counterparts in the decoder. Pix2Pix [9] performs image-to-image translation and style transfer using a U-Net cGAN with an L1 reconstruction constraint between synthesized and target images and a patch-based discriminator (PatchGAN) designed to model the local, high-frequency nuances of the data. While Pix2Pix models local and fine-grained image statistics, perceptual loss [10] has been proposed to measure the difference in the high-level, semantic features of images through the use of a pre-trained, fixed-loss "perceptual network" such as VGG16 [16] pretrained on ImageNet [3]. Different from Pix2Pix, our approach uses a combination of global L2 reconstruction loss and perceptual loss to emphasize the high-level semantic features more likely to be present in both domains over the reconstruction of domain-specific local features.

Converting near-infrared periocular images to visible wavelength images or vice-versa enables the use of intra-spectral recognition algorithms. Hernandez-Diaz *et al.* [8] employ Pix2Pix for this purpose, showing its effectiveness for data augmentation, but ultimately achieve lower performance than a direct feature-learning approach [18]. Alternatively, Taherkhani *et al.* [17] uses coupled cGANs for extreme off-pose face recognition, demonstrating its potential for learning multi-modal representations.

III. Coupled GAN

The overall architecture of the proposed Coupled GAN (CoGAN) is primarily composed of two cGANs. One cGAN processes visible-wavelength images while the other processes NIR images. The networks are "coupled" by the joint task of learning a mapping from periocular images to feature representations in a common latent subspace. While our ultimate goal is to learn discriminative and domain-invariant feature encodings for cross-spectral periocular recognition, we propose to simultaneously optimize for a set of auxiliary objectives related to image reconstruction. We hypothesize these secondary reconstruction tasks can help the feature learning process key into important visual features.

Fig. 2 illustrates the CoGAN's main components. The generators have a U-Net [14] encoder-decoder structure with ResNet-18 [5] encoders (minus the final softmax layer) followed by deconvolutional decoder layers. The discriminators are VGG-like [16] 4-layer CNNs (with filters sizes of 16, 32, 64, and 128, respectively) followed by a fully-connected layer with a single scalar output. The perceptual network is an ImageNet pre-trained VGG16 network. Once the model has been trained, the encoders can operate in isolation from the rest of the CoGAN.

1) Shared Feature Subspace Learning: The ultimate goal of the Coupled GAN is to conduct cross-spectral matching by learning discriminative, domain-invariant features. The contrastive loss [2] term is the portion of the objective function which is most directly related to our ultimate goal. All other losses are auxiliary to the learning process.

Minimizing the contrastive loss encourages the features extracted from a pair of genuine images (i.e. instances of the same class) to be similar and the features of imposter pairs (instances of different classes) to be least some margin m apart. Let $X_V = \{x_V^i\}_{i=1}^N$ and $X_I = \{x_I^j\}_{j=1}^N$ be the visible wavelength and near-infrared periocular images, respectively. Let $\mathcal{D}_z(x_V^i; x_I^j)$ be the distance measure between the features extracted from a pair of images. Given the feature encodings

$z_1(x_V^i)$ and $z_2(x_I^j)$ obtained from the encoder sub-networks, we calculate the distance measure as the L2-norm:

$$\mathcal{D}_z(x_V^i; x_I^j) = \left\| z_1(x_V^i) - z_2(x_I^j) \right\|_2 . \quad (2)$$

Let us define ground-truth labels $y^{ij} = 0$ for genuine pairs and $y^{ij} = 1$ for imposter pairs. The contrastive loss function is then written as:

$$\mathcal{L}_C(x_V^i; x_I^j; y^{ij}) = (1 - y^{ij})\frac{1}{2}(\mathcal{D}_z(x_V^i, x_I^j)^2 + (y^{ij})\frac{1}{2}(\max(0, m - \mathcal{D}_z(x_V^i, x_I^j)))^2, \quad (3)$$

$$\mathcal{L}_C^* = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{L}_C(x_V^i, x_I^j, y^{ij}), \quad (4)$$

where $\mathcal{L}^*$ denotes a combined loss computed over both visible and NIR domains.

A. Image Synthesis

Instead of performing cross-spectral image synthesis, we utilize the pair of cGANs to refine the feature learning process via three auxiliary loss terms: adversarial loss, reconstruction loss, and perceptual loss; all of which work in concert with the contrastive loss to help guide the encoders in learning important visual features.

1) Adversarial Loss: Using (1) for the adversarial loss of a single cGAN, we condition on the input images in order to recreate them. Let G_V and G_I be generators operating on visible and IR images, respectively, while D_V and D_I are their respective discriminators. The total adversarial loss for the coupled networks is:

$$\mathcal{L}_A^* = \frac{1}{N^2}((\sum_{i=1}^{N}\mathcal{L}_{cGAN}(G_V, D_V, x_V^i, x_V^i)) + (\sum_{j=1}^{N}\mathcal{L}_{cGAN}(G_I, D_I, x_I^j, x_I^j))). \quad (5)$$

2) Reconstruction Loss: Minimizing the reconstruction loss trains the GAN to recreate low-level, fine-grained features. We use the L2 distance between the pixel values of a given image x and its synthesized version $G(x)$ as the reconstruction loss. The total reconstruction loss is defined as follows:

$$\mathcal{L}_R(x; G) = \|x - G(x)\|_2^2 , \quad (6)$$

$$\mathcal{L}_R^* = \frac{1}{N^2}(\sum_{i=1}^{N}\mathcal{L}_R(x_V^i, G_V) + (\sum_{j=1}^{N}\mathcal{L}_R(x_I^j, G_I))). \quad (7)$$

3) Perceptual Loss: Given an image x and its reconstruction, the perceptual loss [10] measures the distance between the high-level features extracted by the kth layer of a fixed loss network ϕ. Let $\phi_k(\cdot)$ be the $C_k \times W_k \times H_k$ activations of the kth network layer for a given input image. The perceptual loss is the L2 distance between the feature representations:

$$\mathcal{L}_P(x; G) = \frac{1}{CWH}\sum_{c=1}^{C_k}\sum_{w=1}^{W_k}\sum_{h=1}^{H_k}\|\phi_k(x) - \phi_k(G(x))\|_2^2 , \quad (8)$$

$$\mathcal{L}_P^* = \frac{1}{N^2}(\sum_{i=1}^{N}\mathcal{L}_P(x_V^i, G_V) + \sum_{j=1}^{N}\mathcal{L}_P(x_I^j, G_I)). \quad (9)$$

We use the activations of the last convolutional layer of a pre-trained VGG16 network for ϕ_k.

B. Training and Implementation

The overall objective function of the Coupled GAN is the summation of the multi-task losses plus an additional L2 weight decay term:

$$\mathcal{L} = \lambda_C\mathcal{L}_C^* + \lambda_A\mathcal{L}_A^* + \lambda_R\mathcal{L}_R^* + \lambda_P\mathcal{L}_P^* + \lambda_{L2}\mathcal{L}_{L2}^*, \quad (10)$$

where λ represents the coefficients of the individual loss terms. This function is minimized by Stochastic Gradient Descent via the Adam optimizer.

We set $\lambda_A = \lambda_R = \lambda_P = 1.0$. Optimal values for λ_C vary slightly depending on the dataset (see Section V-B). The learning rate and L2 weight decay (λ_{L2}) are fixed at 1×10^{-4}. Training is done for 300 epochs, with every mini-batch composed of 100 NIR-visible image pairs, resized to 256 × 256, for a total mini-batch size of 200 images. Each pair has a 50% chance of being either a genuine or imposter pair. A 5-fold cross validation scheme on the training data was employed to tune hyperparameters and determine the number of training epochs. The models were implemented with Pytorch 1.7.0 on machines with 2x Tesla V100 GPUs. The code has been made available at https://github.com/vonclites/cogan.

IV. DATASETS

Our approach is benchmarked on the *open world* protocols of Hong Kong PolyU [13] and Cross-Eyed [15] datasets following the method used in [18]. Dataset details are provided in Table I and example images in Fig. 1. Following standard practice in periocular recognition, we consider a subject's left and right eyes to be unique classes. Verification is conducted using a one-against-all pairwise matching strategy.

The Hong Kong PolyU [13] dataset is composed of simultaneously acquired images in the NIR and visible spectrums. The entire database has 12,540 images with a resolution of 640×480. In both visible and thermal spectrums, there are 15 samples of each eye (left and right) from 209 subjects (418 classes). The first 104 subjects plus the left eye of the 105th are assigned to the training set and the rest to testing.

The Cross-Eyed [15] dataset has 3,840 images from 120 subjects (240 classes). There are eight samples from each

TABLE I: Open world protocol of the Hong Kong PolyU and Cross-Eyed cross-spectral periocular datasets.

Dataset	Train/Test Subjects	Train/Test Classes	Train/Test Images	Gen/Imposter Test Pairs
HK PolyU	$104.5/104.5$	$209/209$	$6,270/6,270$	$21,945/9,781,200$
Cross-Eyed	$60/60$	$120/120$	$1,920/1,920$	$3,360/913,920$

TABLE II: Comparison with State-of-the-Art and baseline.

Dataset	Model	AUC%	EER%	FRR@FAR	
				1%	10%
HK PolyU	ResNet-50	96.03	9.93	31.22	11.25
	Zanlorensi [18]	-	8.02	-	-
	ResNet-18	98.16	5.85	18.6	4.36
	CoGAN	**98.65**	**5.14**	**12.32**	**3.27**
Cross-Eyed	ResNet-50	95.43	10.00	26.43	10.04
	Zanlorensi [18]	-	4.39	-	-
	ResNet-18	98.95	4.15	10.19	2.62
	CoGAN	**99.41**	**3.07**	**6.11**	**1.66**

of the classes for each spectrum. All images are 400×300 resolution and were obtained at a distance of 1.5 meters. The first 60 subjects are assigned to the training set and the remaining 60 to the test set.

Images are rescaled to 256×256 with zero mean and unit variance per training data statistics. For data augmentation, training images are zero-padded to 272×272 and a 256×256 crop is extracted.

V. Results and Discussion

In this section, we present and discuss the benchmark results of the CoGAN framework along with additional experiments and ablation studies investigating the benefits of our hybrid approach. Once trained, the only components which are utilized for evaluating the CoGAN's performance are the encoder sub-networks. Performance is evaluated in terms of the Area Under the ROC Curve (AUC) and Equal Error Rate (EER). Also provided is the False Rejection Rate at False Accept Rates of 1% and 10% (FRR@FAR=1%, FRR@FAR=10%).

A. Comparison with State of the Art

At the time of writing, the only recent work which reports results on the *open world* protocols of the benchmark datasets is, to the best of our knowledge, Zanlorensi *et al.* [18], who have achieved state-of-the-art performance on the cross-spectral scenario. In Table II, the performance of the proposed CoGAN model is compared with the state-of-the-art and two baselines models. The baseline versions are equivalent to training the CoGAN using only contrastive loss and weight decay. The CoGAN achieves improved EER(%) over the ResNet18 baseline and the larger, coupled ResNet-50 model of [18], which in turn performs better than our ResNet-50 baseline, potentially due to their use of cosine similarity. ResNet-18 can also be seen to outperform ResNet-50, consistent with the intra-spectral periocular recognition results reported in [11].

TABLE III: Feature Dimension Size $|z|$ vs. Contrastive Cost Coefficient λ_C in terms of AUC(%). λ_A, λ_R, λ_P set to 1.0.

Dataset	$\lvert z\rvert$	λ_C			
		1.0	2.0	**5.0**	10.0
HK PolyU	32	97.91	97.85	98.16	97.89
	64	98.35	98.21	**98.65**	98.42
	128	98.25	97.95	97.87	97.92
	256	98.28	98.54	98.09	98.38
Cross-Eyed	32	98.69	98.74	98.94	98.88
	64	98.94	98.84	98.58	98.94
	128	99.01	99.24	99.31	**99.41**
	256	98.08	97.63	97.88	98.03

TABLE IV: The effect of the auxiliary image synthesis tasks (adversarial loss λ_A, reconstruction loss λ_R, and perceptual loss λ_P) on the HK PolyU dataset.

λ_C	λ_A	λ_R	λ_P	AUC, %
1.0	0.0	0.0	0.0	98.41
5.0	1.0	0.0	0.0	98.54
5.0	1.0	1.0	0.0	98.55
5.0	1.0	0.0	1.0	98.63
5.0	1.0	1.0	1.0	**98.65**

B. Ablation Studies

The following ablation studies serve to further evaluate the performance of the CoGAN and the effect of the auxiliary image synthesis tasks on the end-goal of periocular verification. Table III lists the AUC(%) for various sizes of the feature encoding vectors (z_1 and z_2) and contrastive cost coefficients (λ_C). Similar to [18], we also found the optimal feature size to be smaller for the Hong Kong dataset than the Cross-Eyed.

Table IV shows the gradual improvements gained by incorporating the additional loss terms of the GAN. The first row of the table represents a baseline version of the model constrained only by the primary contrastive loss objective. We can see that the largest impact over the baseline is made by a combination of adversarial and perceptual loss. This suggests that by emphasizing the importance of capturing the high-level, global appearance, the CoGAN is better able to guide the learning of discriminative features.

VI. Conclusion

We show our novel approach to cross-spectral periocular recognition achieves state-of-the-art results using the proposed CoGAN architecture. Our experiments demonstrate that by introducing auxiliary intra-spectral image reconstruction tasks to support the effort of shared subspace feature learning for cross-spectral periocular recognition, the CoGAN attains higher performance over a baseline version of the model. Our work presents the CoGAN architecture as a promising framework for further research in both intra-spectral and cross-spectral periocular recognition, given the compact nature of its ResNet-18 backbone, lack of need for well-aligned image pairs, and potential for cross-spectral synthesis applications. In particular, our investigation reveals that perceptual loss may aid in disentangling domain-specific information from the feature representations used for matching, indicating a potential direction for future work.

References

[1] S. S. Behera, B. Mandal, and N. B. Puhan. Twin deep convolutional neural network-based cross-spectral periocular recognition. In *2020 National Conference on Communications (NCC)*, pages 1–6. IEEE, 2020.

[2] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] K. Hernandez-Diaz, F. Alonso-Fernandez, and J. Bigun. Periocular recognition using cnn features off-the-shelf. In *2018 International conference of the biometrics special interest group (BIOSIG)*, pages 1–5. IEEE, 2018.

[7] K. Hernandez-Diaz, F. Alonso-Fernandez, and J. Bigun. Cross spectral periocular matching using resnet features. In *2019 International Conference on Biometrics (ICB)*, pages 1–7. IEEE, 2019.

[8] K. Hernandez-Diaz, F. Alonso-Fernandez, and J. Bigun. Cross-spectral periocular recognition with conditional adversarial networks. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020.

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

[10] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[11] P. Kumari and K. Seeja. Periocular biometrics for non-ideal images: with off-the-shelf deep cnn & transfer learning approach. *Procedia Computer Science*, 167:344–352, 2020.

[12] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[13] P. R. Nalla and A. Kumar. Toward more accurate iris recognition using cross-spectral matching. *IEEE transactions on Image processing*, 26(1):208–221, 2016.

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

[15] A. Sequeira, L. Chen, P. Wild, J. Ferryman, F. Alonso-Fernandez, K. B. Raja, R. Raghavendra, C. Busch, and J. Bigun. Cross-eyed-cross-spectral iris/periocular recognition database and competition. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2016.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] F. Taherkhani, V. Talreja, J. Dawson, M. C. Valenti, and N. M. Nasrabadi. Pf-cpgan: Profile to frontal coupled gan for face recognition in the wild. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020.

[18] L. A. Zanlorensi, D. R. Lucio, A. d. S. B. Junior, H. Proença, and D. Menotti. Deep representations for cross-spectral ocular biometrics. *IET Biometrics*, 9(2):68–77, 2019.