

Exploring Complementarity and Explainability in CNNs for Periocular Verification Across Acquisition Distances

Fernando Alonso-Fernandez

School of Information Technology
Halmstad University, Sweden
feralo@hh.se

Kevin Hernandez Diaz

School of Information Technology
Halmstad University, Sweden
kevin.hernandez-diaz@hh.se

Jose M. Buades

Computer Graphics and Vision and AI Group
University of Balearic Islands, Spain
josemaria.buades@uib.es

Kiran Raja

Faculty of Information Technology and Electrical Engineering
NTNU, Norway
kiran.raja@ntnu.no

Josef Bigun

School of Information Technology
Halmstad University, Sweden
josef.bigun@hh.se

Abstract—We study the complementarity of different CNNs for periocular verification at different distances on the UBIPr database. We train three architectures of increasing complexity (SqueezeNet, MobileNetv2, and ResNet50) on a large set of eye crops from VGGFace2. We analyse performance with cosine and χ^2 metrics, compare different network initialisations, and apply score-level fusion via logistic regression. In addition, we use LIME heatmaps and Jensen–Shannon divergence to compare attention patterns of the CNNs. While ResNet50 consistently performs best individually, the fusion provides substantial gains, especially when combining all three networks. Heatmaps show that networks usually focus on distinct regions of a given image, which explains their complementarity. Our method significantly outperforms previous works on UBIPr, achieving a new state-of-the-art.

Index Terms—Periocular biometrics, CNN fusion, score-level fusion, Explainable AI, Jensen–Shannon divergence, LIME

I. INTRODUCTION

The periocular region (area around the eye) is a robust biometric trait, especially under unconstrained or degraded conditions where full-face or iris capture may not be viable [1]. Compared to face or iris, it offers a balanced trade-off between accuracy, acquisition ease, and robustness to occlusion and resolution changes. Partial face visibility is also common in contexts such as social media [2], masks, work gear, cultural coverings, etc. [3]. Although convolutional neural networks (CNNs) dominate feature learning in biometrics [4], their use in periocular recognition remains limited [3], [5], [6], partly due to the lack of large dedicated datasets [5].

Several studies used off-the-shelf deep features for periocular recognition (including fusion) by leveraging networks pre-trained on ImageNet [7]–[11] or face datasets [7], [9], [12]. However, they are not specifically trained for periocular. Other works [13]–[22] trained CNNs on small/mid periocular sets like UFPR (33k images) [15], VISOB 2.0 (158k) [23], UBIPr (3.3k) [24], or synthetic data [19]. This contrasts

with face recognition, which benefits from massive sets (e.g., WebFace260M [25]). Despite progress, most studies rely on a single architecture and do not examine network complementarity. The impact of acquisition distance is also underexplored, as is attention analysis across networks, especially in multi-network setups.

Accordingly, we analyse three CNNs of varying complexity (SqueezeNet, MobileNetv2, ResNet50) for periocular recognition. To overcome scarcity of large training sets, we use >1.9M ocular crops from VGGFace2 [26], and evaluate on UBIPr [24] using intra- and inter-distance protocols. Our CNN comparison across distances shows that ResNet50 leads individually, whereas score-level fusion via logistic regression leads to consistent improvements, especially when combining all networks. The use of LIME heatmaps [27] and Jensen–Shannon divergence to visualise and quantify attention patterns reveals a clear complementarity among networks. Our results show the importance of architectural diversity in enhancing performance and set a new state-of-the-art on the UBIPr dataset.

II. RECOGNITION NETWORKS

We evaluate three CNNs of increasing complexity: SqueezeNet [28] (light, 18 layers, 1.24M params), MobileNetv2 [29] (medium, 53, 3.5M), and ResNet50 [30] (large, 50, 25.6M). ResNet50 uses residual blocks that improve gradient flow and support deeper architectures. Each block reduces dimensionality with 1×1 convolutions, applies 3×3 filters in the reduced space, and restores the original size. MobileNetv2 uses inverted residuals and depth-wise separable convolutions to reduce parameters and inference time. Shortcut connections link thinner layers instead, with intermediate representations in a higher-dimensional space. SqueezeNet is a compact, non-residual network that first ‘squeezes’ dimensionality with 1×1 filters, then ‘expands’ them using 1×1 and

3×3 convolutions in a lower-dimensional space. This selection enables comparison of networks of varying complexity. We adapt the ImageNet-pretrained models in MATLAB R2024b by changing the first convolution stride from 2 to 1, allowing 113×113 inputs. Images are normalised by subtracting 127.5 and dividing by 128. For SqueezeNet, we follow [31], adding batch norm between convolutions and ReLUs, missing in the original implementation.

III. DATASETS

We train using VGGFace2 (3.31M images, 9,131 identities) [26] (Figure 1), which contains significant variation in pose, age, lighting, and background. Using dataset annotations, we crop ocular regions from the 8,631 training identities (3.14M images). Images are aligned (eye centres horizontal), scaled to 113 pixels inter-eye distance, and cropped into two 113×113 patches centred on each eye. We ensure both eyes are visible by requiring the horizontal eye-centre distance and nose vertical to be within 40% of the inter-eye distance, and discard faces below 50 px inter-eye to avoid strong upsampling. Left eye crops are flipped for orientation consistency, and both eyes are treated as the same identity, resulting in 953,786 valid faces and 1,907,572 ocular crops (221 per identity on average).

Testing uses the UBIPr periocular database [24], with images from 4–8 m captured by a CANON EOS 5D. We select 1,718 frontal images (86 subjects with two sessions, one image per eye/session/distance, totalling $86 \times 2 \times 2 = 344$ images per distance). Images are manually annotated for inner/outer eye boundaries, resized to match the average sclera radius R_s of their distance group, and aligned by cropping a $7.6R_s \times 7.6R_s$ square around the sclera centre (Figure 1). Likewise, left eye crops are flipped, and both eyes represent the same identity. Sclera boundary is used for normalisation due to its stability across dilation and better contrast than the pupil boundary. Images are finally resized to match CNN input. Unlike methods that mask the iris [32], we retain the full periocular region to simulate realistic conditions where iris segmentation is unreliable (e.g., low resolution, blur, or low pigmentation in visible images) [1].

In some cases, networks are pretrained for face recognition. Following [31], ImageNet-initialised models are first trained on RetinaFace-cleaned MS1M [33] (5.1M images, 93.4K classes), then fine-tuned on VGGFace2 (3.14M images). As before, left eye crops are flipped, and both eyes treated as the same subject. This two-step training leverages MS1M large volume and VGGFace2 greater intra-class diversity, having demonstrated superior performance [26], [31].

IV. TRAINING AND RECOGNITION PROTOCOLS

The networks are trained for ocular identification using cross-entropy loss on VGG2 crops. We use SGDM (batch=128, learning rate=0.01, 0.005, 0.001, 0.0001, reduced when validation plateaus). Validation uses 2% of training images per user. Model weights are initialised from either ImageNet or pretrained face recognition models. Verification

is done on the UBIPr dataset for intra- and inter(cross)-distance scenarios. Identity templates are extracted from the layer before classification (Global Average Pooling). Images at distance D_i are compared against those at D_j , with $i, j \in \{1, 2, 3, 4, 5\}$. Intra-distance ($i = j$) genuine scores are obtained by comparing eyes of session 1 vs eyes of session 2 of the same subject (4 comparisons/user), giving $86 \times 4 = 344$ scores per distance. Cross-distance ($i \neq j$) scores compare eyes of session 1 at D_i vs eyes of both sessions at D_j (8 comparisons/user), giving $86 \times 8 = 688$ scores per inter-distance pair. Impostor scores are obtained by comparing eyes from session 1 of a user to eyes from session 2 of all other users, giving $86 \times 85 \times 4 = 29240$ impostor scores per distance combination. This results in 8600 genuine (5 intra + 10 cross) and 438600 impostor scores across 15 distance combinations. As comparison metrics, we use cosine similarity, commonly used in CNN-based verification, and χ^2 distance, which has also demonstrated strong performance in related works [7].

We also apply score-level fusion via linear logistic regression to combine multiple networks. Given N networks producing scores $(s_{1j}, s_{2j}, \dots, s_{Nj})$ for trial j , the fused score is $f_j = a_0 + a_1 \cdot s_{1j} + a_2 \cdot s_{2j} + \dots + a_N \cdot s_{Nj}$, with weights $a_0, a_1, \dots, a_N$ trained via logistic regression [34], [35]. This approach outperforms simple fusion rules (mean, sum) [36] and common classifiers in multibiometrics like SVM or Random Forest [9]. It achieved top results in ocular benchmarks through expert fusion [37] and recent works [9], [11] involving both traditional and off-the-shelf CNNs. Though it is a weighted sum, the coefficients are optimised with a discriminative learning rule [38].

V. RESULTS

5.1 Individual Networks

We begin by presenting (Table I, top) ocular verification results on the UBIPr database for the three networks, jointly considering intra- and inter-distance scores. The χ^2 distance consistently provides better performance than cosine similarity, confirming earlier findings [7]. While the improvement is marginal in some cases, it exceeds 0.5% for SqueezeNet, the weakest network. Regarding initialisation, the best case is always ImageNet. Although one might expect that starting from face-pretrained networks would be advantageous, since the networks are familiar with eye regions, our results suggest otherwise. Face models may be overly specialised to full-face features, whereas ImageNet models begin with more primitive, generic features, which can better adapt to ocular data. This supports the established view of ImageNet as a robust, versatile foundation for downstream tasks [40], including biometrics [7]–[11], [41]. In absolute performance, residual networks (MobileNetv2, ResNet50) outperform SqueezeNet, with ResNet50, the largest one, achieving the best EER at 1.66%.

5.2 Network Combination

We then conduct (also Table I, top) fusion of the networks. Consistent with previous observations, both χ^2 distance and ImageNet initialization remain the most effective choices.

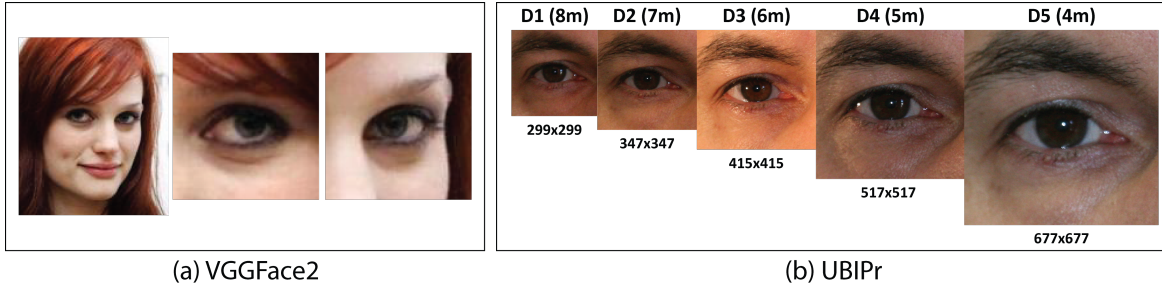


Fig. 1. Example images from the databases employed. The relative scale differences among normalised UBIPr images are shown, as well as their resulting size.

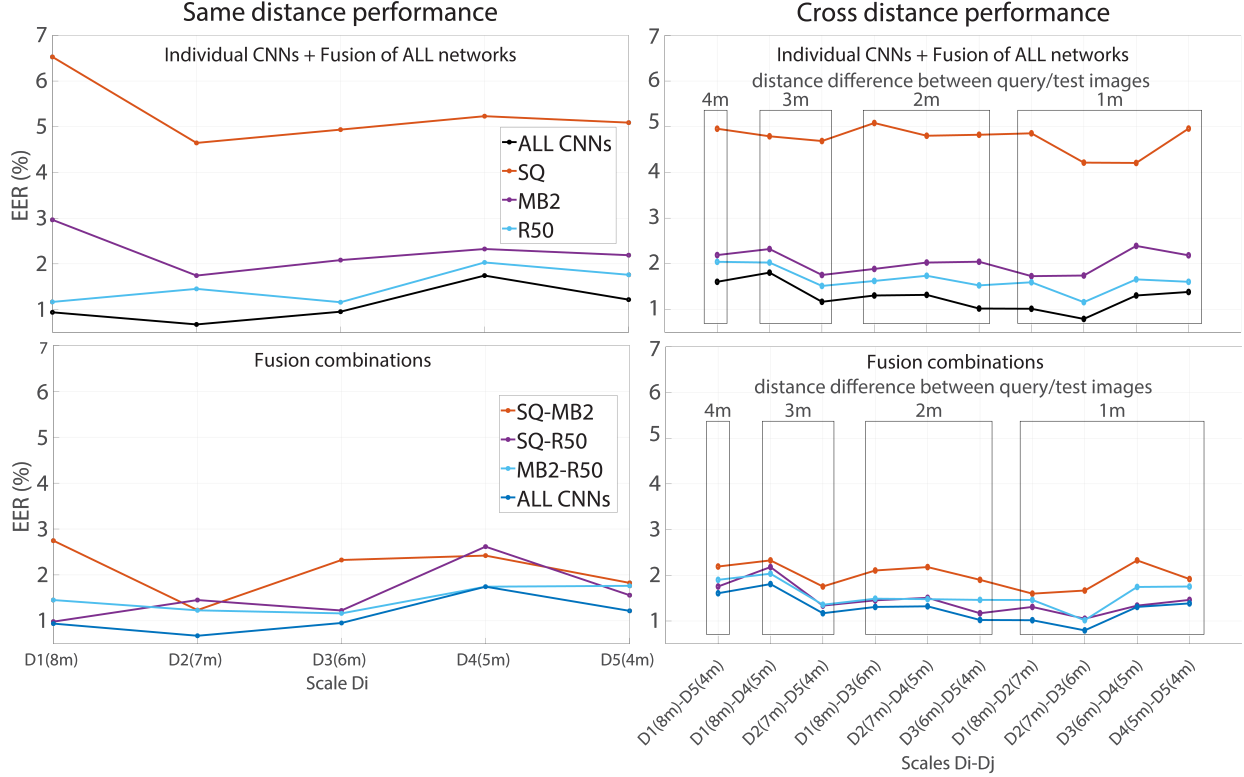


Fig. 2. Ocular verification results (EER %) on UBIPr for scale variation experiments (ImageNet initialization, χ^2 distance). The figure shows the performance of the individual networks (top) and of the different fusion combinations (bottom). The top plot also shows the fusion of all networks (best fusion case) for comparison with the individual networks.

Notably, combining two CNNs improves performance, with MobileNetv2 + ResNet50 generally being the best case. This indicates strong complementarity between these architectures, as also seen in face recognition [27]. Despite both being residual networks, their differing residual layer structures likely promote diverse learned features, thus complementarity. In contrast, fusions involving the simpler, less accurate SqueezeNet lead to smaller gains. Only SqueezeNet with the powerful ResNet50 occasionally approaches the performance of MobileNetv2 + ResNet50. Interestingly, fusing all three networks achieves the best performance, with improvements exceeding 20%, compared to the 7–10% gain when fusing just two models. These findings highlight the advantage of leveraging architectural diversity rather than relying solely

on individual model strength. Even lower-performing models like SqueezeNet can provide complementary information that enhances the system by compensating for the limitations of stronger models [42], [43].

5.3 Comparison with Previous Works

We also report as reference in Table I (bottom) previous works on UBIPr. Direct comparisons should be made with caution, as differences in experimental protocols occur, evidenced by the number of images used or scores reported in these works. A key difference in our study, shared only with [11], is the alignment of eye crops by flipping to a common orientation and the same identity. This increases the number of genuine comparisons while making impostor comparisons more challenging by removing anatomical asymmetry bias. Despite

network	Initialization / Metric							
	ImageNet				Face			
	cosine		χ^2		cosine		χ^2	
SQ	5.44	-	4.93	-	5.97	-	5.45	-
MB2	2.12	-	2.10	-	2.24	-	2.15	-
R50	1.73	-	1.66	-	1.95	-	1.93	-
SQ+MB2	2.05	(-3.07%)	2.04	(-2.80%)	2.13	(-4.80%)	2.05	(-4.35%)
SQ+R50	1.61	(-6.96%)	1.51	(-9.37%)	1.81	(-7.16%)	1.81	(-6.56%)
MB2+R50	1.61	(-7.27%)	1.59	(-4.67%)	1.77	(-8.96%)	1.74	(-10.17%)
ALL	1.33	(-23.40%)	1.31	(-21.14%)	1.49	(-23.20%)	1.50	(-22.22%)

Results of other works of the literature

Work	Feature type	Features	EER
[24]	traditional	SIFT+LBP+HOG	16%
[39]	traditional	SIFT+LBP+HOG	8.4%
	traditional	SIFT+LBP+HOG+SAFE	7.9%
[7]	traditional	SIFT+LBP+HOG	9.1%
	deep	CNN (ResNet101)	5.6%
	trad. + deep	SIFT+LBP+HOG + CNN (ResNet101)	5.1%
[14]	deep	AttNet + FCN-Peri	2.26%
[11]	trad.	SIFT+LBP+HOG	10.58%
	deep	CNN off-the-shelf (ResNet50)	8.53%
	deep	ViT off-the-shelf (tiny)	11.48%
	deep	CNN ots (ResNet50) + ViT ots (base)	7.72%
	trad. + deep	SIFT+LBP+HOG + CNN (ResNet50) + ViT (base)	6.32%
This work	deep	SqueezeNet + MobileNetv2 + ResNet50	1.31%

TABLE I

OCULAR VERIFICATION RESULTS (EER %) ON UBIPr FOR DIFFERENT NETWORK INITIALISATIONS (ImageNet AND FACE RECOGNITION WEIGHTS) AND COMPARISON METRICS (COSINE SIMILARITY AND χ^2 DISTANCE). THE RELATIVE EER VARIATION IN FUSION EXPERIMENTS WITH RESPECT TO THE BEST INDIVIDUAL NETWORK IS GIVEN IN BRACKETS. THE TABLE ALSO SHOWS RESULTS FROM PREVIOUS WORKS ON THE SAME DATABASE.

this more challenging evaluation, our results outperform all previous works. The seminal UBIPr paper [24] set the initial benchmark, improved later by others with handcrafted features such as SIFT, LBP, HOG, or SAFE. More recent works employ deep embeddings from off-the-shelf CNNs and ViTs pretrained on ImageNet as generic feature extractors. In contrast, we fine-tune several networks on >1.9M eye crops from the large VGGFace2 dataset, achieving state-of-the-art performance on UBIPr.

5.4 Distance Variation

Based on the previous results, we adopt χ^2 distance and ImageNet initialisation for the rest of the experiments. We then analyse varying acquisition distances. Figure 2 (left column) shows the EER when both images are captured at the same distance (intra-distance), ordered from farthest (left in the x-axes) to closest (right). The right column presents the inter-distance case, grouping results by the distance gap between image pairs, from 4 meters (left in the x-axes) to 1 meter (right). From Figure 2 (left column), we observe that performance is generally stable across intra-distances, except at the farthest point (8m). ResNet50 consistently performs best, with EERs

below 2%, while SqueezeNet performs worst. MobileNetv2 stays below 3% across all ranges. In the inter-distance setting (right column), performance degrades with increasing distance differences, especially for SqueezeNet. ResNet50 remains the most robust, with EERs under 2% even at a 4m gap. Fusion results confirm that combining all CNNs consistently gives the best accuracy, achieving EER<1.5%, and in some cases <1%, across most distance scenarios. Fusing any two networks also improves performance, though less effectively than using all three.

5.5 Explainability Analysis

To further explore complementarity between networks, we analyse LIME heatmaps [27], which highlight the most relevant pixels for each model. To quantify similarity between heatmaps, we use the Jensen–Shannon divergence (JSD), a symmetric, smoothed version of the Kullback–Leibler (KL) divergence commonly used to compare probability distributions P and Q . It is defined as $JSD(P \parallel Q) = 0.5 \cdot KL(P \parallel M) + 0.5 \cdot KL(Q \parallel M)$, where $M = 0.5 \cdot (P + Q)$ is the average distribution, and $KL(P \parallel Q) = \sum_i P(i) \log(P(i)/Q(i))$ is the standard KL divergence. Heatmaps are normalized into prob-

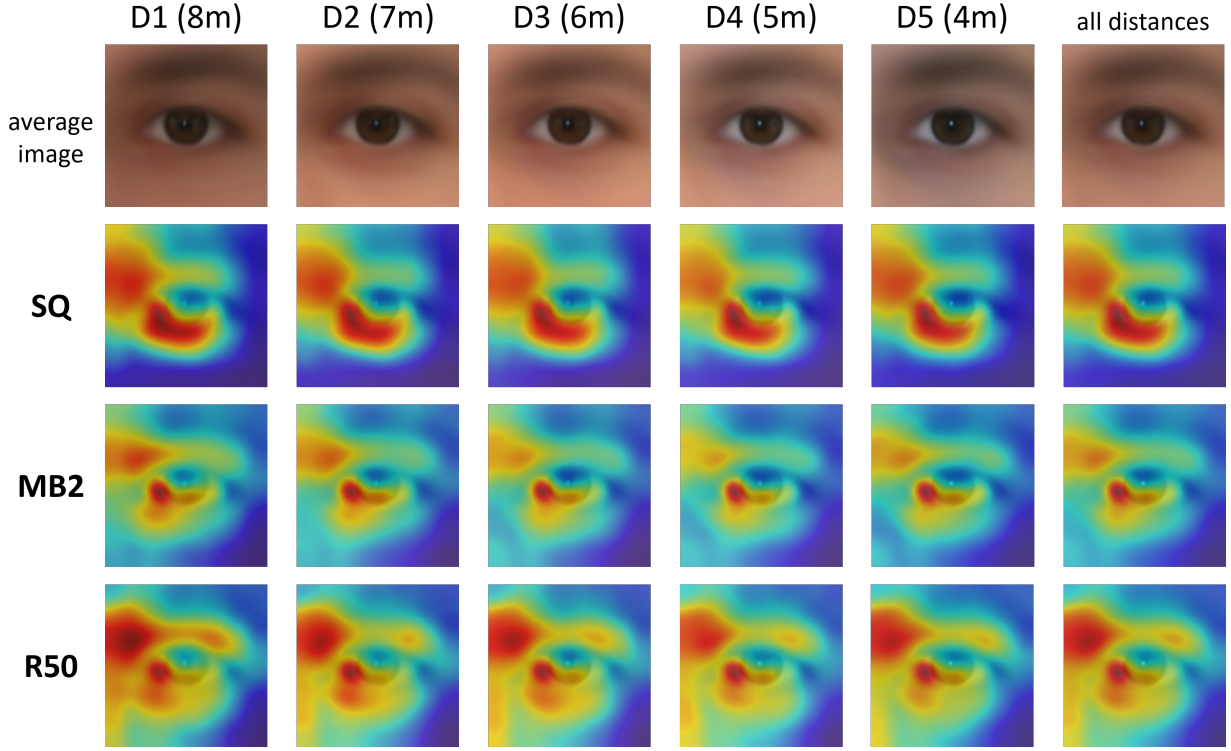


Fig. 3. Average LIME heatmaps on UBIPr per distance (columns) and CNN (rows).

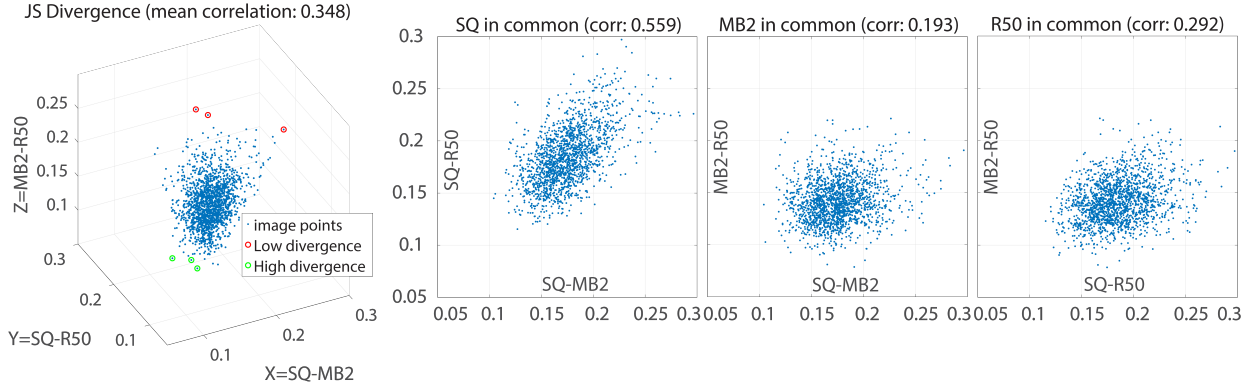


Fig. 4. Jensen-Shannon divergence between the heatmaps generated by the networks. The 3D scatter plot on the left represents divergence values across images for each pair of CNNs. The three plots on the right show the 2D projections onto each pair of axes. Correlation values are also given. The clouds are computed with images of the entire database (all distances).

ability distributions by dividing each pixel by the total sum. JSD ranges from 0 (identical) to $\log(2) \approx 0.6931$ (maximally different distributions). Figure 3 presents the average LIME heatmaps of each network at different distances. Recall that left eyes are horizontally flipped for orientation consistency (nose on the left). Overall, MobileNetv2 has more localised and compact activations, particularly under the lower eyelid, while ResNet50 and SqueezeNet show broader patterns. All models highlight regions like the upper eyelid, sclera, and tear duct, especially ResNet50 and SqueezeNet. The cheek and right periocular part receive minimal attention, and interestingly, the pupil/iris is also less attended, suggesting reliance

on periocular context rather than iris texture. Across distances, the heatmaps remain relatively stable within each network, in line with the consistent performance seen in Figure 2 (left).

While average heatmaps reveal common attended regions, we also assess differences per image. To do so, we compute the pairwise JSD between networks for each image and plot the results in a 3D scatter space (Figure 4), where each axis represents JSD for a specific network pair. The cloud shape suggests low correlation between divergence values, indicating that the networks often produce complementary explanations. In particular, 2D projections onto SQ-MB2 vs. MB2-R50 and SQ-R50 vs. MB2-R50 planes are near-circular, with

low Pearson correlations (0.193, 0.292). A more linear trend appears in SQ-MB2 vs. SQ-R50, with a moderate correlation of 0.559, suggesting that SqueezeNet tends to agree/disagree similarly with the other two CNNs. However, the overall lack of strong correlation supports the benefit of fusing networks with distinct attention patterns. Figure 5 shows examples with the lowest and highest average JSD between network pairs (marked in Figure 4). Interestingly, the lowest-divergence cases often involve glasses, which the networks learn to ignore, focusing on consistent periocular features such as the skin below the lower eyelid, tear duct, and sclera. Regarding the highest-divergence cases, they show varied attention to the sclera, tear duct, lower eyelid, or eyelashes.

VI. CONCLUSIONS

This work analysed the performance and complementarity of three CNNs of different complexity and depth for periocular verification under scale variation on the UBIPr database [24]. We observed that deeper, residual networks (e.g., ResNet50) perform best individually, but the best results are obtained when combining all three models. Score-level fusion using logistic regression provided up to 23% relative improvement over the best network. Using LIME-based heatmaps [27] and Jensen–Shannon divergence, we further showed that each network focuses on different regions of the eye, suggesting that their feature representations are complementary. This explains the success of the fusion strategy and support the use of explainability tools to guide architectural decisions. Our method establishes new state-of-the-art results on the UBIPr dataset and demonstrates the value of combining diverse CNN architectures for robust periocular verification.

Despite being captured at several distances, UBIPr contains high-resolution images given by a CANON EOS 5D camera (22.3 MPx) and cooperative subjects. It remains to be seen how well the system generalizes to more challenging scenarios, including lower-quality sensors such as those used in surveillance environments, and non-cooperative subjects. We would also like to test our approach in near-infrared data, for which spectrum translation techniques may be employed [44] to ensure sufficient training data in this domain as well. We are also working on integrating more discriminative loss functions, such as margin-based approaches like ArcFace. Another avenue is the adoption of a sequential fine-tuning strategy, where networks are first trained on ocular crops from MS-Celeb-1M (MS1M) and later refined using VGGFace2. This approach can exploit the larger scale of MS1M for initial generalization and benefit from the greater intra-class variability in VGGFace2, an strategy seen very effective strategy in face recognition [26], [31].

Acknowledgements. This work was partly done while F. A.-F. was visiting researcher at University of Balearic Islands. F. A.-F., K. H.-D., and J. B. thank the Swedish Research Council (VR) and the EU (HORIZON Europe project PopEye under Grant Agreement no 101168317). Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the

European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work is part of the Project PID2022-136779OB-C32 (PLEISAR) funded by MICIU/ AEI /10.13039/501100011033/ and FEDER, EU.

REFERENCES

- [1] F. Alonso, J. Bigun, J. Fierrez, N. Damer, H. Proenca, and A. Ross, “Periocular biometrics: A modality for unconstrained scenarios,” *Computer*, vol. 57, no. 06, pp. 40–49, jun 2024.
- [2] P. Hedman, V. Skepetzis, K. Hernandez, J. Bigun, and F. Alonso, “On the effect of selfie beautification filters on face detection and recognition,” *Patt Recogn Lett*, 2022.
- [3] R. Sharma and A. Ross, “Periocular biometrics and its relevance to partially masked faces: A survey,” *Computer Vision and Image Understanding*, vol. 226, p. 103583, 2023.
- [4] K. Sundararajan and D. L. Woodard, “Deep learning for biometrics: A survey,” *ACM Comput. Surv.*, vol. 51, no. 3, 2018.
- [5] L. A. Zanolensi, R. Laroca, E. Luz, A. S. B. Jr., L. S. Oliveira, and D. Menotti, “Ocular recognition databases and competitions: A survey,” *Artificial Intelligence Review*, vol. 55, pp. 129–180, 2022.
- [6] D. Zeng, R. Veldhuis, and L. Spreuwers, “A survey of face recognition techniques under occlusion,” *IET Biometrics*, vol. 10, no. 6, pp. 581–606, 2021.
- [7] K. Hernandez, F. Alonso, and J. Bigun, “Periocular recognition using CNN features off-the-shelf,” in *Proc BIOSIG*, 2018.
- [8] —, “Cross spectral periocular matching using resnet features,” in *Proc ICB*, 2019.
- [9] F. Alonso, K. B. Raja, R. Raghavendra, C. Busch, J. Bigün, R. Vera-Rodríguez, and J. Fierrez, “Cross-sensor periocular biometrics in a global pandemic: Comparative benchmark and novel multialgorithmic approach,” *Information Fusion*, 2022.
- [10] K. Hernandez, F. Alonso, and J. Bigun, “One-shot learning for periocular recogn.: Exploring domain adapt. & data bias on deep representations,” *IEEE Access*, vol. 11, 2023.
- [11] F. Alonso, K. Hernandez, P. Tiwari, and J. Bigun, “Combined cnn and vit features off-the-shelf: Another astounding baseline for recognition,” in *Proc. WIFS*, 2024.
- [12] J. E. Tapia, A. Valenzuela, R. Lara, M. Gomez-Barrero, and C. Busch, “Selfie periocular verification using an efficient super-resolution approach,” *IEEE Access*, vol. 10, pp. 67 573–67 589, 2022.
- [13] Z. Zhao and A. Kumar, “Accurate periocular recognition under less constrained environment using semantics-assisted conv. neural network,” *IEEE TIFS*, vol. 12, no. 5, 2017.
- [14] —, “Improving periocular recognition by explicit attention to critical regions in deep neural network,” *IEEE TIFS*, vol. 13, no. 12, pp. 2937–2952, 2018.
- [15] L. A. Zanolensi, R. Laroca, D. R. Lucio, L. R. Santos, A. S. B. Jr., and D. Menotti, “A new periocular dataset collected by mobile devices in unconstrained scenarios,” *Scientific Reports*, vol. 12, p. 17989, 2022.
- [16] J. N. Kolf, F. Boutros, F. Kirchbuchner, and N. Damer, “Lightweight periocular recognition through low-bit quantization,” in *Proc. IJCB*, 2022.
- [17] V. Talreja, N. M. Nasrabadi, and M. C. Valenti, “Attribute-based deep periocular recogn.: Leveraging soft biometrics to improve periocular recogn.” in *Proc. WACV*, 2022.
- [18] A. Almadan and A. Rattani, “Benchmarking neural network compression techniques for ocular-based user authentication on smartphones,” *IEEE Access*, vol. 11, 2023.
- [19] J. N. Kolf, J. E., F. Boutros, H. Proença, and N. Damer, “Syper: Synthetic periocular data for quantized light-weight recognition in the nir and visible domains,” *Image and Vis. Comp.*, vol. 135, p. 104692, 2023.
- [20] J. N. Kolf, J. Elliesen, N. Damer, and F. Boutros, “Mixquantbio: Towards extreme face and periocular recognition model compression with mixed-precision quantization,” *Eng Appl of AI*, 2024.
- [21] P. Coelho, G. Silva, D. Oliveira, G. Moreira, E. Luz, and P. Silva, “Periocular efficientnet: A deep model for periocular recognition,” in *Proc. LA-CCI*, 2024.
- [22] L. G. Fonseca Carreira, D. Menotti, and W. R. Schwartz, “Dpr-v2s: A deep framework for periocular recognition in surveillance environments,” in *Proc. SIBGRAPI*, 2024.

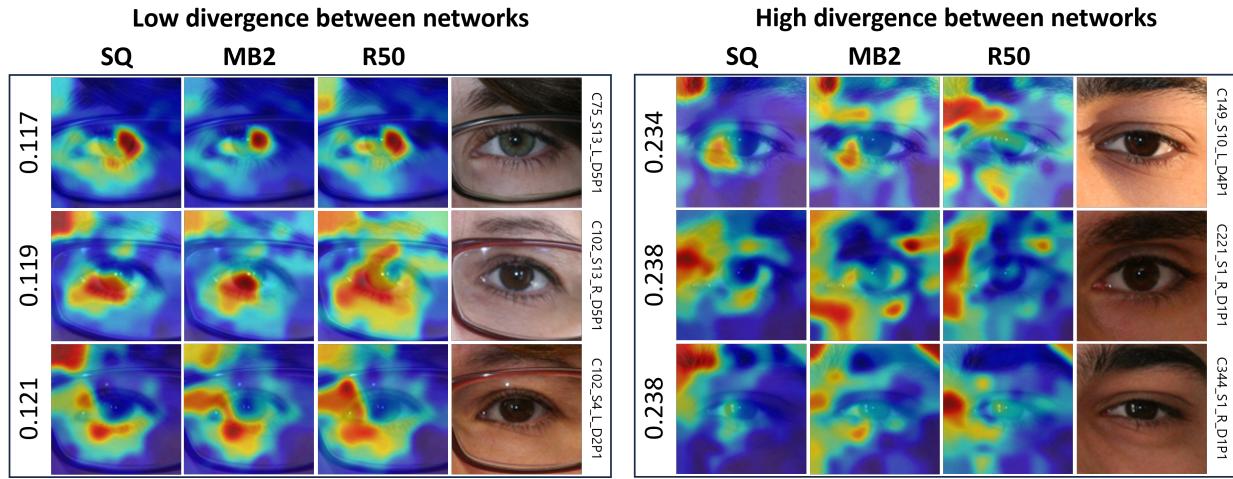


Fig. 5. Individual heatmaps where the three networks diverge the least (left) and the most (right). The numeric values indicate the average JS divergence between the three possible network pairs.

- [23] H. Nguyen, N. Reddy, A. Rattani, and R. Derakhshani, "Visob 2.0 - the second international competition on mobile ocular biometric recognition," in *Proc. ICPRW*, 2021.
- [24] C. Padole and H. Proenca, "Periocular recognition: Analysis of performance degradation factors," *Proc ICB*, 2012.
- [25] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, D. Du, J. Lu, and J. Zhou, "Webface260m: A benchmark for million-scale deep face recognition," *IEEE TPAMI*, vol. 45, no. 2, pp. 2627–2644, 2023.
- [26] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *Proc. FG*, 2018.
- [27] F. Alonso, K. Hernandez, J. M. Buades, P. Tiwari, and J. Bigun, "An explainable model-agnostic algorithm for cnn-based biometrics verification," in *Proc. WIFS*, 2023.
- [28] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size," *CoRR*, vol. abs/1602.07360, 2016. [Online]. Available: <http://arxiv.org/abs/1602.07360>
- [29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc CVPR*, 2018.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc CVPR*, 2016.
- [31] F. Alonso, J. Barrachina, K. H. Diaz, and J. Bigun, "Squeezefaceposenet: Lightweight face verification across different poses for mobile platforms," in *Proc. WMWB-ICPR*, 2020.
- [32] U. Park, R. R. Jillela, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum," *IEEE TIFS*, vol. 6, no. 1, 2011.
- [33] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset & benchmark for large-scale face recognition," in *Proc. ECCV*, 2016.
- [34] S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the NIST'99 1-speaker submissions," *Digital Signal Processing*, vol. 10, pp. 237–248, 2000.
- [35] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwartz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006," *IEEE TASSP*, 2007.
- [36] F. Alonso, J. Fierrez, D. Ramos, and J. Ortega-Garcia, "Dealing With Sensor Interoperability in Multi-biometrics," *Proc. SPIE-BTHI*, 2008.
- [37] A. F. Sequeira, L. Chen, J. Ferryman, F. Alonso, J. Bigun, K. B. Raja, R. Raghavendra, C. Busch, and P. Wild, "Cross-eyed - cross-spectral iris/periocular recognition database and competition," in *Proc BIOSIG*, 2016.
- [38] E. Bigun, J. Bigun, B. Duc, and S. Fischer, "Expert Conciliation for Multi Modal Person Authentication Systems by Bayesian Statistics," *Proc AVBPA*, vol. Springer LNCS-1206, pp. 291–300, 1997.
- [39] F. Alonso, A. Mikaelian, and J. Bigun, "Compact multi-scale periocular recognition using SAFE features," in *Proc. ICPR*, 2016.
- [40] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *Proc CVPRW*, 2014.
- [41] K. Nguyen, C. Fookes, A. Ross, and S. Sridharan, "Iris recognition with off-the-shelf cnn features: A deep learning perspective," *IEEE Access*, vol. 6, pp. 18 848–18 855, 2018.
- [42] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics. part 1: Fundamentals and review," *Information Fusion*, vol. 44, pp. 57–64, 2018.
- [43] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Information Fusion*, vol. 52, pp. 187 – 205, 2019.
- [44] K. Hernandez, F. Alonso, and J. Bigun, "Cross-spectral periocular recognition with conditional adversarial networks," in *Proc. IJCB*, 2020.