PIC-Score: Probabilistic Interpretable Comparison Score for Optimal Matching Confidence in Single- and Multi-Biometric Face Recognition

Pedro C. Neto[1,2], Ana F. Sequeira[1,2], Jaime S. Cardoso[1,2] and Philipp Terhörst[3]
[1]INESC TEC, Porto, Portugal
[2]Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
[3]Paderborn University, Paderborn, Germany

1. Supplementary

In the following, we will provide more information on the face recognition systems and confidence estimation methods used in the experiments. This aims to make the experiments more comprehensible.

1.1. Face Recognition Systems

To show that the used FRS work properly, Figure 1 shows the ROC curves for each FRS on the different databases. The performance of these SOTA systems is highly accurate. Moreover, these systems do barely any errors on the less challenging LFW [3] and Morph [6] databases. The same can be concluded from the first row of Figures 2, 3, 4, 5 and 6. There, the comparison score distributions are shown for all face recognition systems. The genuine and imposter distributions are highly separable on LFW and Morph. On the other side, the genuine and imposter distributions have less separability on Adience [2] and ColorFeret [8]. Thus, for some FRS there is a visible overlap between a portion of the samples in both distributions.

1.2. Confidence Estimation Methods

To get a better understanding of the confidence estimation methods from previous works, Figures 2, 3, 4, 5 and 6 show confidence score distributions of all baseline methods, PIC scores and comparison score distributions. Each figure has these scores shown for a specific FRS. It is visible that independently of the FRS used, only PIC and ERBC [5] show a probabilistic interpretation of the confidence. Nonetheless, the probabilistic interpretation of ERBC is not as insightful as PIC interpretation as seen in Figure 3i and 3e. In this previous example, the ERBC confidence for impostor images is spread across the entire range of values.

Furthermore, the x-axis of LRC-based [12] confidence score distributions indicates a wide range of confidence values. This removes any capability of interpreting the con-fidence values. Moreover, despite the wide range, the majority of the values lies between larger numbers which, as seen in Figure 3m, might impact the separability. DTC [4] score distribution strongly resembles the comparison scores distribution. Besides not having a probabilistic interpretation, the range of confidence values varies with the scores produced by each FRS, as seen in Figures 6q and 5q.

These plot demonstrate that most methods do not possess a probabilistic interpretation. Whilst PIC shows a strong separability for every dataset and FRS tested. Moreover, the values of PIC do not vary across datasets and FRS, which potentiates its usage as a tool to effectively verify and inform a human operator of low-confidence samples.

References

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019. 2

[2] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Trans. Information Forensics and Security*, 9(12):2170–2179, 2014. 1

[3] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 1

[4] Marco Huber, Philipp Terhörst, Florian Kirchbuchner, Naser Damer, and Arjan Kuijper. Stating comparison score uncertainty and verification decision confidence towards transparent face recognition. In *33nd British Machine Vision Conference 2022, BMVC 2022, November 21-24, 2022*. BMVA Press, 2022. 1

[5] Marco Huber, Philipp Terhörst, Anh Thi Luu, Florian Kirchbuchner, and Naser Damer. Verification of sitter identity across historical portrait paintings by confidence-aware face recognition. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022. 1

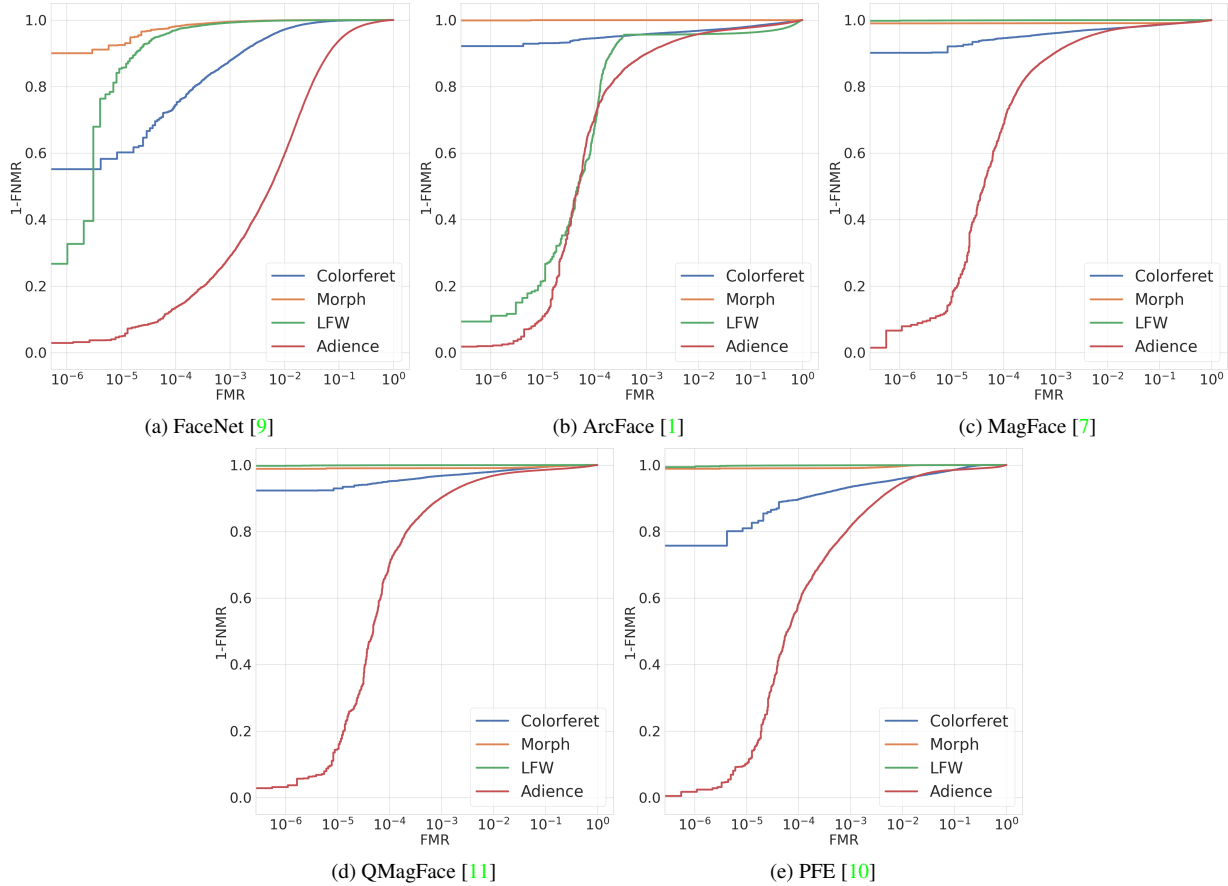| (a) FaceNet [9] | (b) ArcFace [1] | (c) MagFace [7] |

| (d) QMagFace [11] | (e) PFE [10] |

Figure 1. **Performance Face Recognition Systems** - The verification performances for all utilized FRS is shown on the four databases. The systems perform highly accurate, especially on LFW and Morph.

[6] Karl Ricanek Jr. and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FGR 2006), 10-12 April 2006, Southampton, UK*, pages 341–345. IEEE Computer Society, 2006. 1

[7] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14225–14234. Computer Vision Foundation / IEEE, 2021. 2

[8] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, 2000. 1

[9] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society, 2015. 2

[10] Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6901–6910. IEEE, 2019. 2

[11] Philipp Terhörst, Malte Ihlefeld, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran B. Raja, and Arjan Kuijper. Qmagface: Simple and accurate quality-aware face recognition. *CoRR*, abs/2111.13475, 2021. 2

[12] Chris G. Zeinstra, Didier Meuwly, Raymond N. J. Veldhuis, and Luuk J. Spreeuwers. Mind the gap: A practical framework for classifiers in a forensic context. In *9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018*, pages 1–9. IEEE, 2018. 1
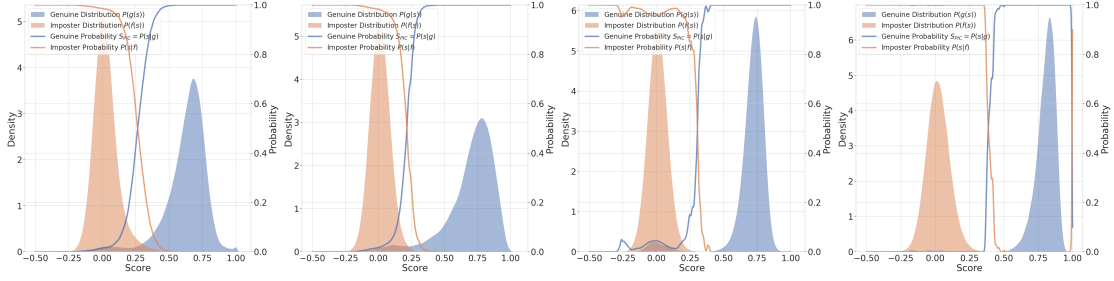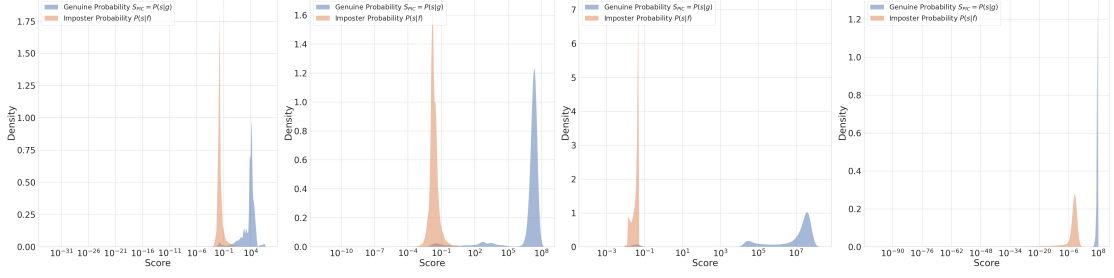
(a) Adience - Comparison Scores

(b) ColorFeret - Comparison Scores

(c) LFW - Comparison Scores

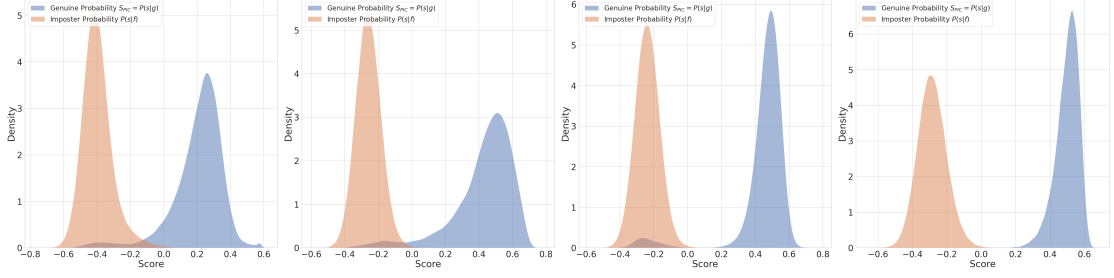(d) Morph - Comparison Scores
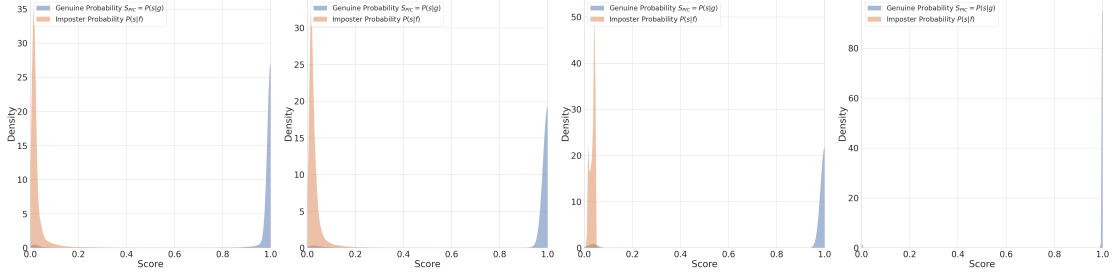
(e) Adience - PIC (ours)

(f) ColorFeret - PIC (ours)

(g) LFW - PIC (ours)

(h) Morph - PIC (ours)
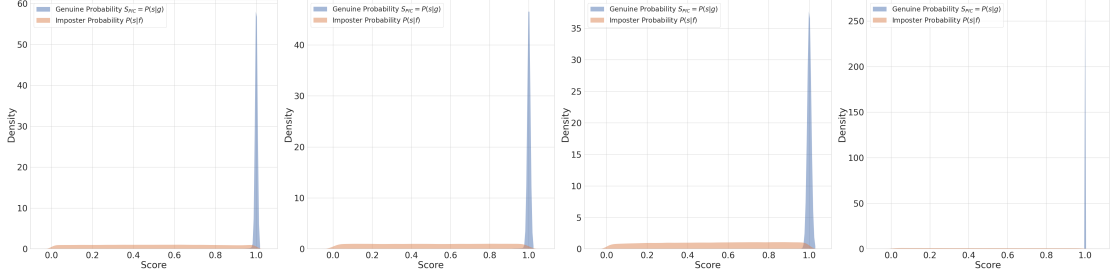
(i) Adience - ERBC

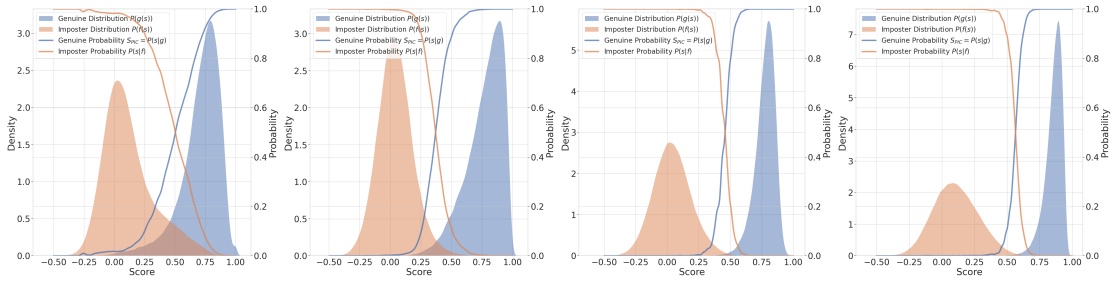(j) ColorFeret - ERBC

(k) LFW - ERBC

(l) Morph - ERBC

(m) Adience - LRC

(n) ColorFeret - LRC
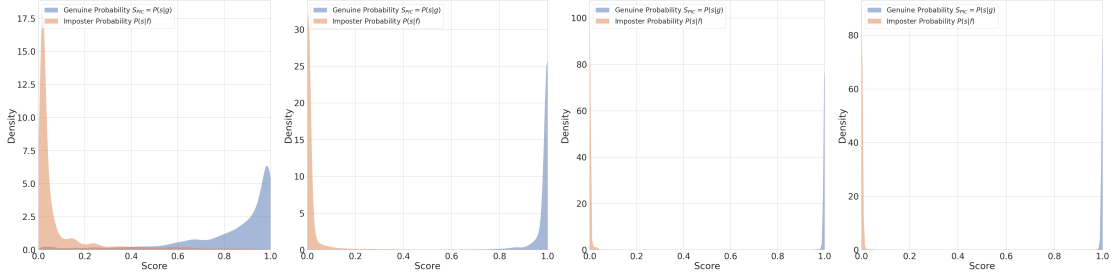
(o) LFW - LRC

(p) Morph - LRC

(q) Adience - DTC
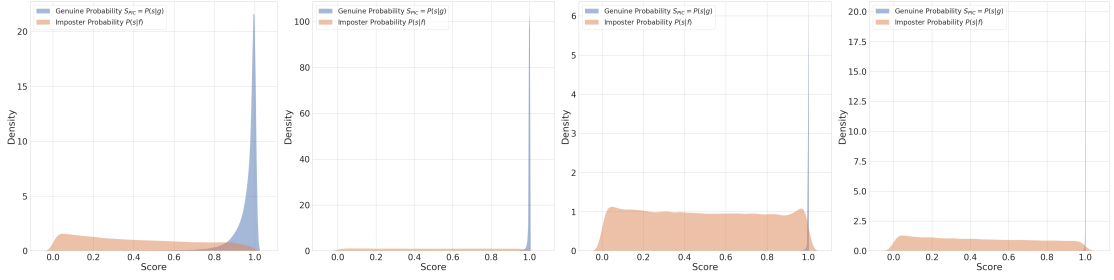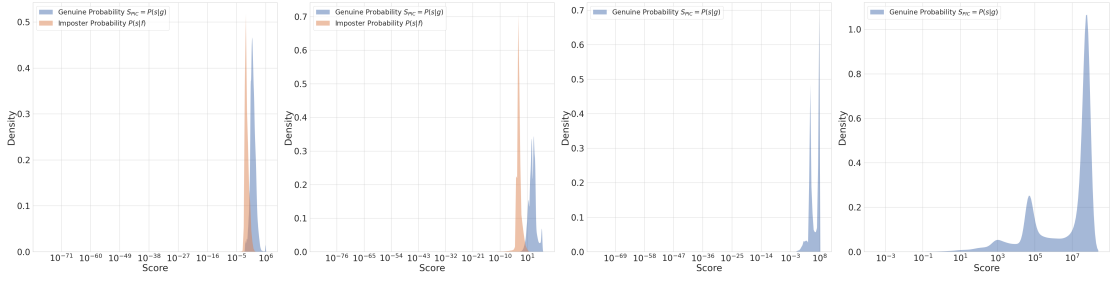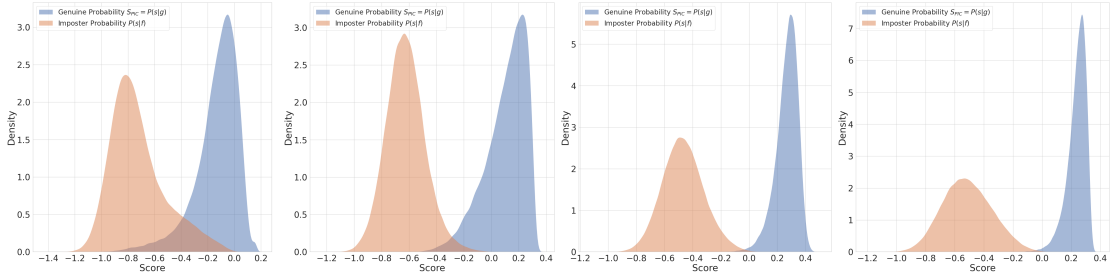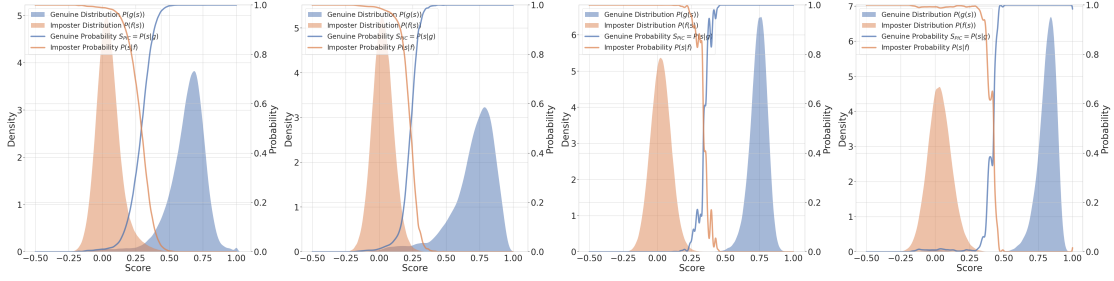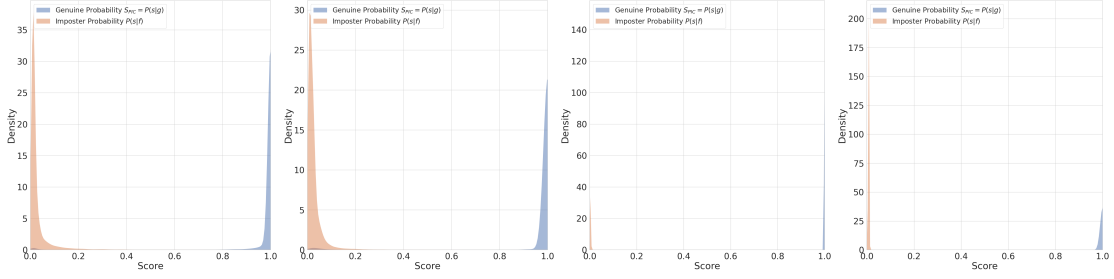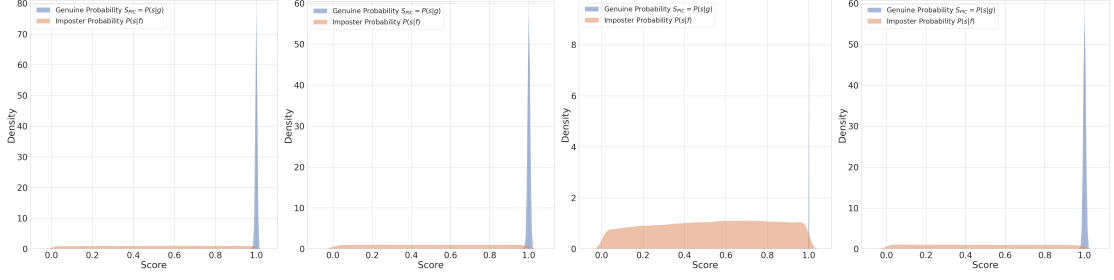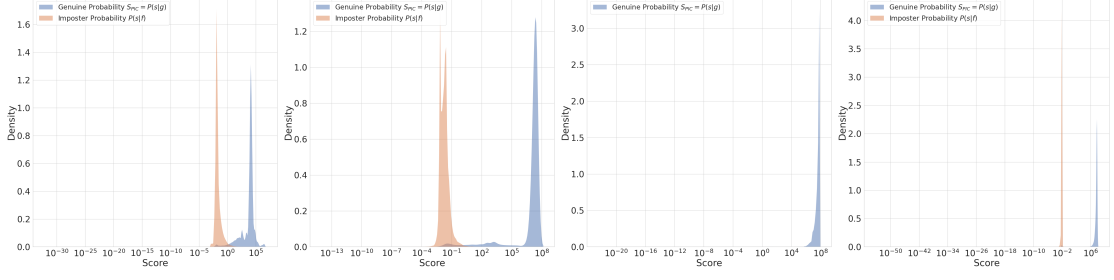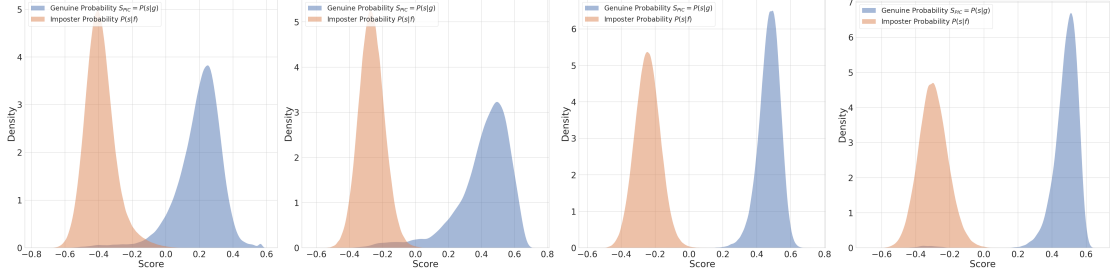
(r) ColorFeret - DTC

(s) LFW - DTC

(t) Morph - DTC

Figure 2. **Score Distribution Analysis** - The original (first row), the PIC (second row), the ERBC (third row), the LRC (fourth row) and the DTC (fifth row) score distributions are shown for ArcFace. Based on the original score distributions, also the corresponding probabilities for genuine and imposter are shown at the top.

(a) Adience - Comparison Scores

(b) ColorFeret - Comparison Scores

(c) LFW - Comparison Scores

(d) Morph - Comparison Scores

(e) Adience - PIC (ours)

(f) ColorFeret - PIC (ours)

(g) LFW - PIC (ours)

(h) Morph - PIC (ours)

(i) Adience - ERBC

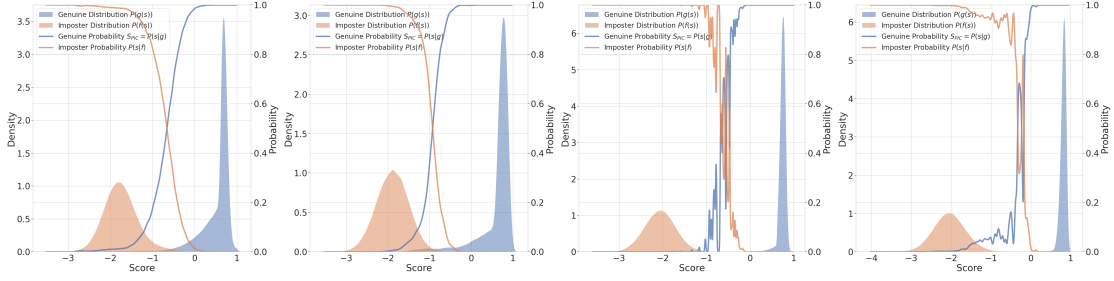(j) ColorFeret - ERBC

(k) LFW - ERBC

(l) Morph - ERBC

(m) Adience - LRC

(n) ColorFeret - LRC

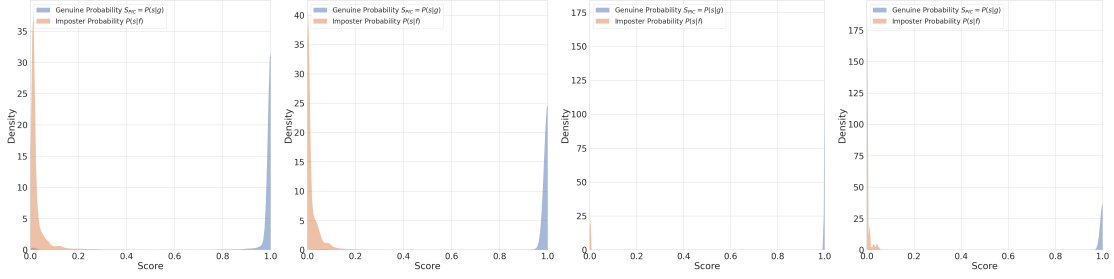(o) LFW - LRC

(p) Morph - LRC

(q) Adience - DTC

(r) ColorFeret - DTC
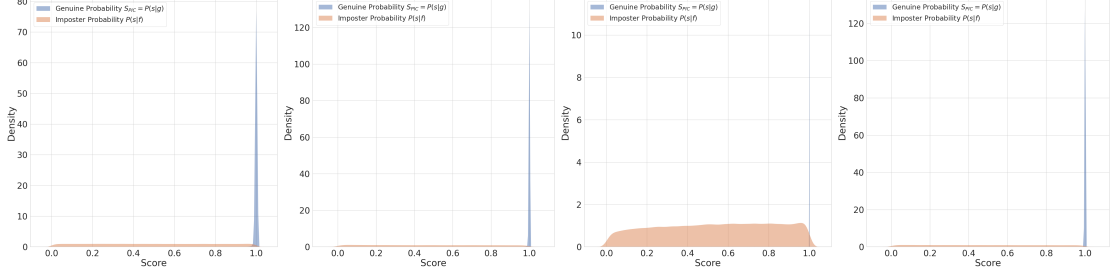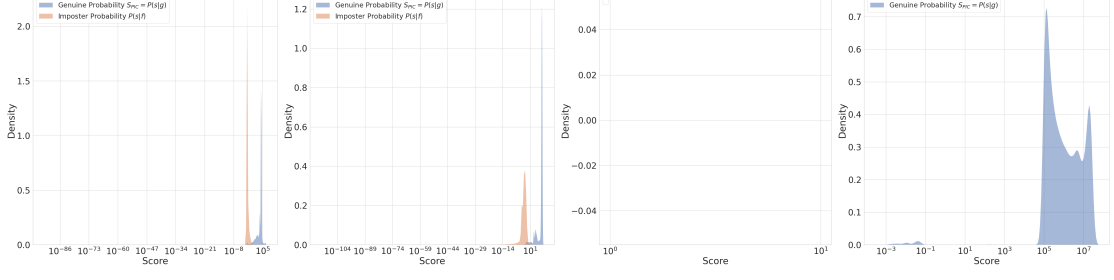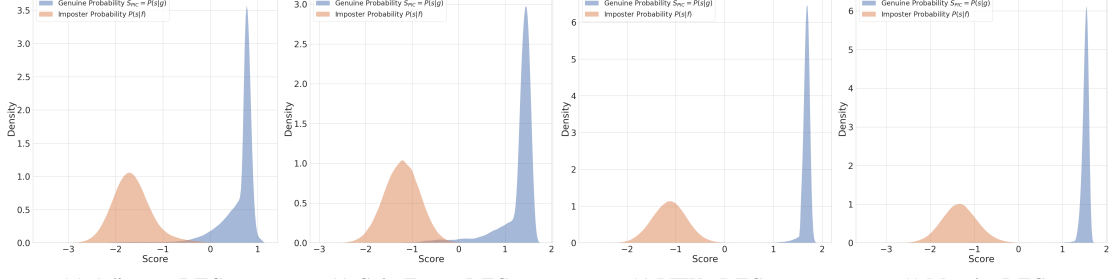
(s) LFW - DTC

(t) Morph - DTC

Figure 3. **Score Distribution Analysis** - The original (first row), the PIC (second row), the ERBC (third row), the LRC (fourth row) and the DTC (fifth row) score distributions are shown for FaceNet. Based on the original score distributions, also the corresponding probabilities for genuine and imposter are shown at the top.

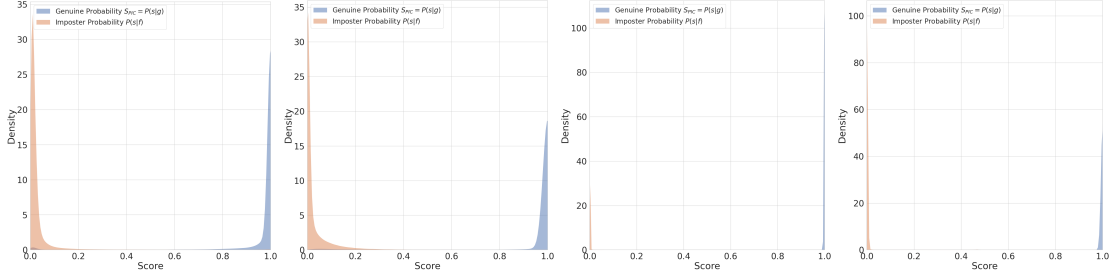Figure 4. **Score Distribution Analysis** - The original (first row), the PIC (second row), the ERBC (third row), the LRC (fourth row) and the DTC (fifth row) score distributions are shown for MagFace. Based on the original score distributions, also the corresponding probabilities for genuine and imposter are shown at the top.
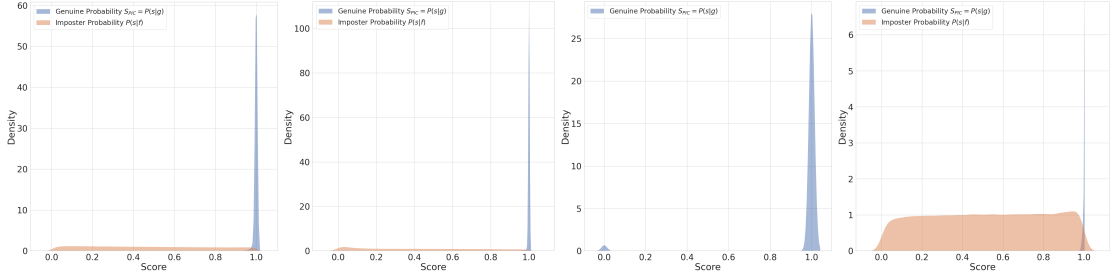
Figure 5. **Score Distribution Analysis** - The original (first row), the PIC (second row), the ERBC (third row), the LRC (fourth row) and the DTC (fifth row) score distributions are shown for QMagFace. Based on the original score distributions, also the corresponding probabilities for genuine and imposter are shown at the top.
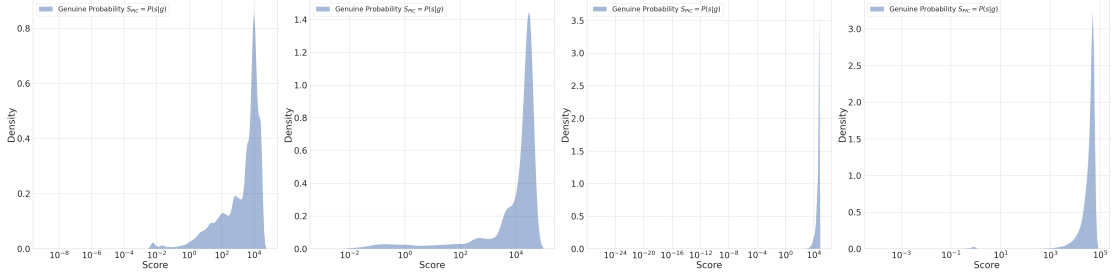
(a) Adience - Comparison Scores

(b) ColorFeret - Comparison Scores

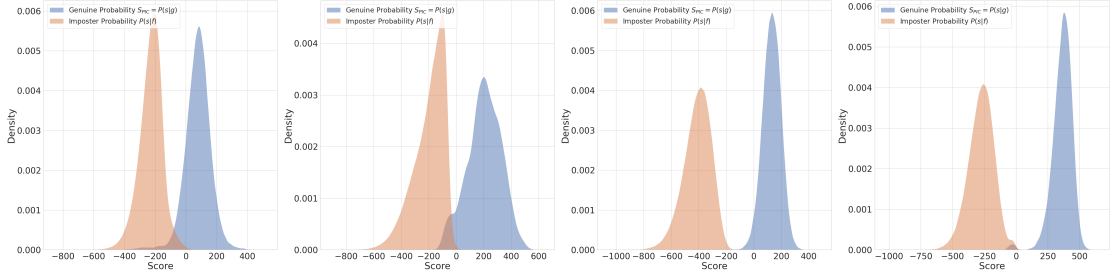(c) LFW - Comparison Scores

(d) Morph - Comparison Scores

(e) Adience - PIC (ours)

(f) ColorFeret - PIC (ours)

(g) LFW - PIC (ours)

(h) Morph - PIC (ours)

(i) Adience - ERBC

(j) ColorFeret - ERBC

(k) LFW - ERBC

(l) Morph - ERBC

(m) Adience - LRC

(n) ColorFeret - LRC

(o) LFW - LRC

(p) Morph - LRC

(q) Adience - DTC

(r) ColorFeret - DTC

(s) LFW - DTC

(t) Morph - DTC

Figure 6. **Score Distribution Analysis** - The original (first row), the PIC (second row), the ERBC (third row), the LRC (fourth row) and the DTC (fifth row) score distributions are shown for PFE. Based on the original score distributions, also the corresponding probabilities for genuine and imposter are shown at the top.