Are you In or Out (of gallery)? Wisdom from the Same-Identity Crowd Supplementary Material

Aman Bhatta[1] Maria Dhakal[1] Michael C. King[2] Kevin W. Bowyer[1*]

[1]University of Notre Dame
[2]Florida Insitute of Technology
{abhatta,kwb}@nd.edu,michaelking@fit.edu

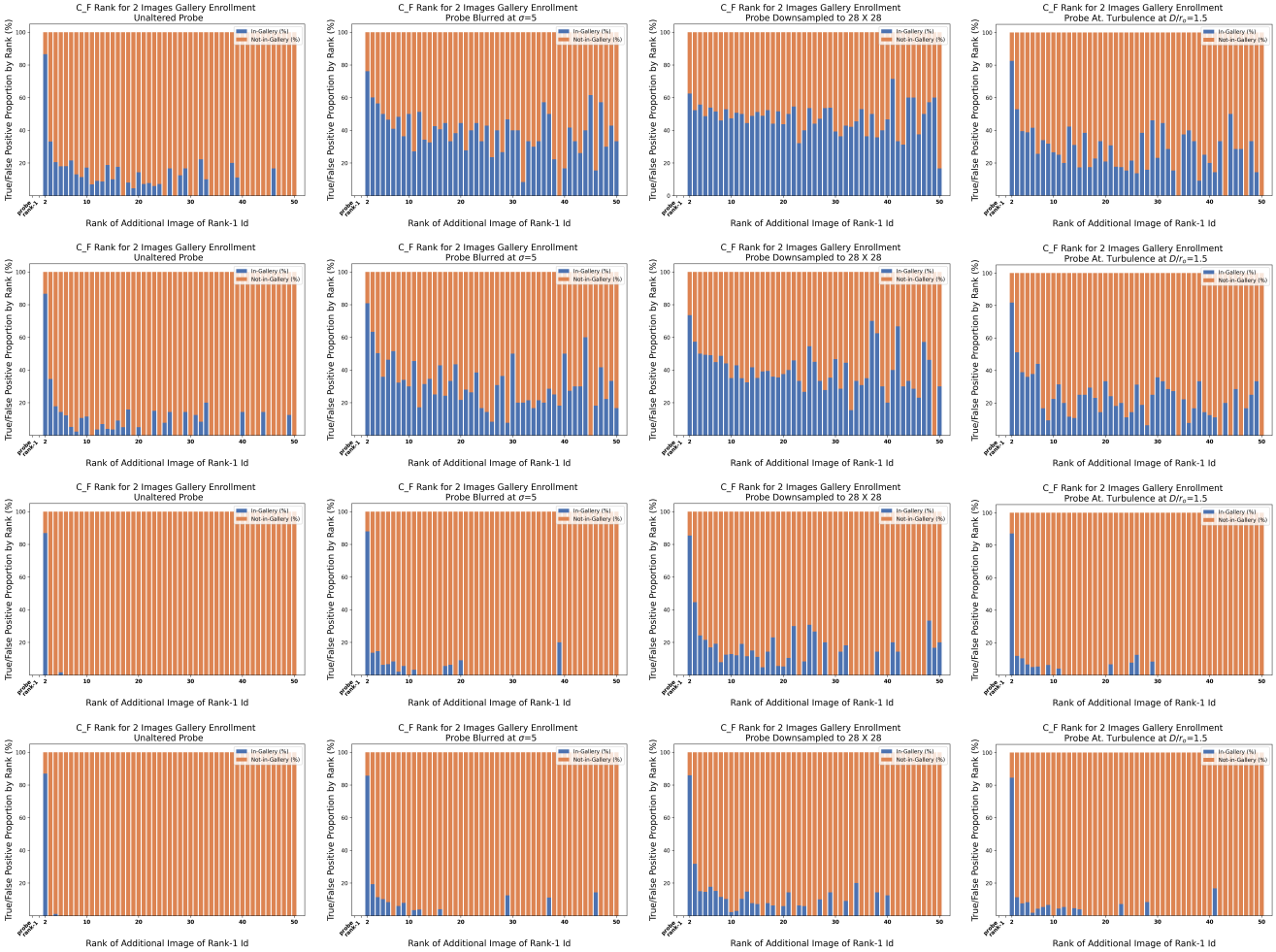
1. Extended Results on Rank Distribution of Additional Images for Rank-One Identity

In the main paper, we presented the rank patterns for an additional image of the rank-one identity on the Caucasian Female cohort using FaceNet and AdaFace loss. Here, we extend these findings to all demographic groups in the MORPH dataset—Caucasian Male/Female and African-American Male/Female—using the four face embedding networks detailed in the main paper, namely FaceNet, ArcFace, TransFace, and AdaFace. The following observations summarize the results:

- For mug-shot quality probes, the rank pattern for the additional image of the rank-one identity displays a clear separation between in-gallery and out-of-gallery samples. This leads to robust recognition performance for all embedding networks.
- Under blurred and atmospheric turbulence conditions, this distinction in rank patterns diminishes noticeably for FaceNet, while remaining moderately intact for ArcFace and largely preserved for AdaFace and TransFace. As a result, FaceNet exhibits a marked drop in accuracy, ArcFace sees a moderate reduction, and AdaFace and TransFace are only minimally affected.
- The most substantial degradation occurs for downsampled probes, where the rank pattern disruption is present across all methods. For FaceNet, this disruption is so pronounced that the pattern appears nearly arbitrary, resulting in near-random predictions for the in-gallery vs. out-of-gallery task. ArcFace also degrades considerably but fares better than FaceNet, whereas AdaFace and TransFace—benefiting from more quality-aware training—retain the overall pattern but still incur some performance loss in downstream in-gallery vs. out-of-gallery classification task.
- Finally, across all demographic groups, performance remains relatively consistent, indicating no significant deviations in rank patterns for any specific subset of the MORPH dataset.

*Dr. Bowyer is a member of the FaceTec (facetec.com) Advisory Board. Results in this paper do not necessarily relate to FaceTec products.

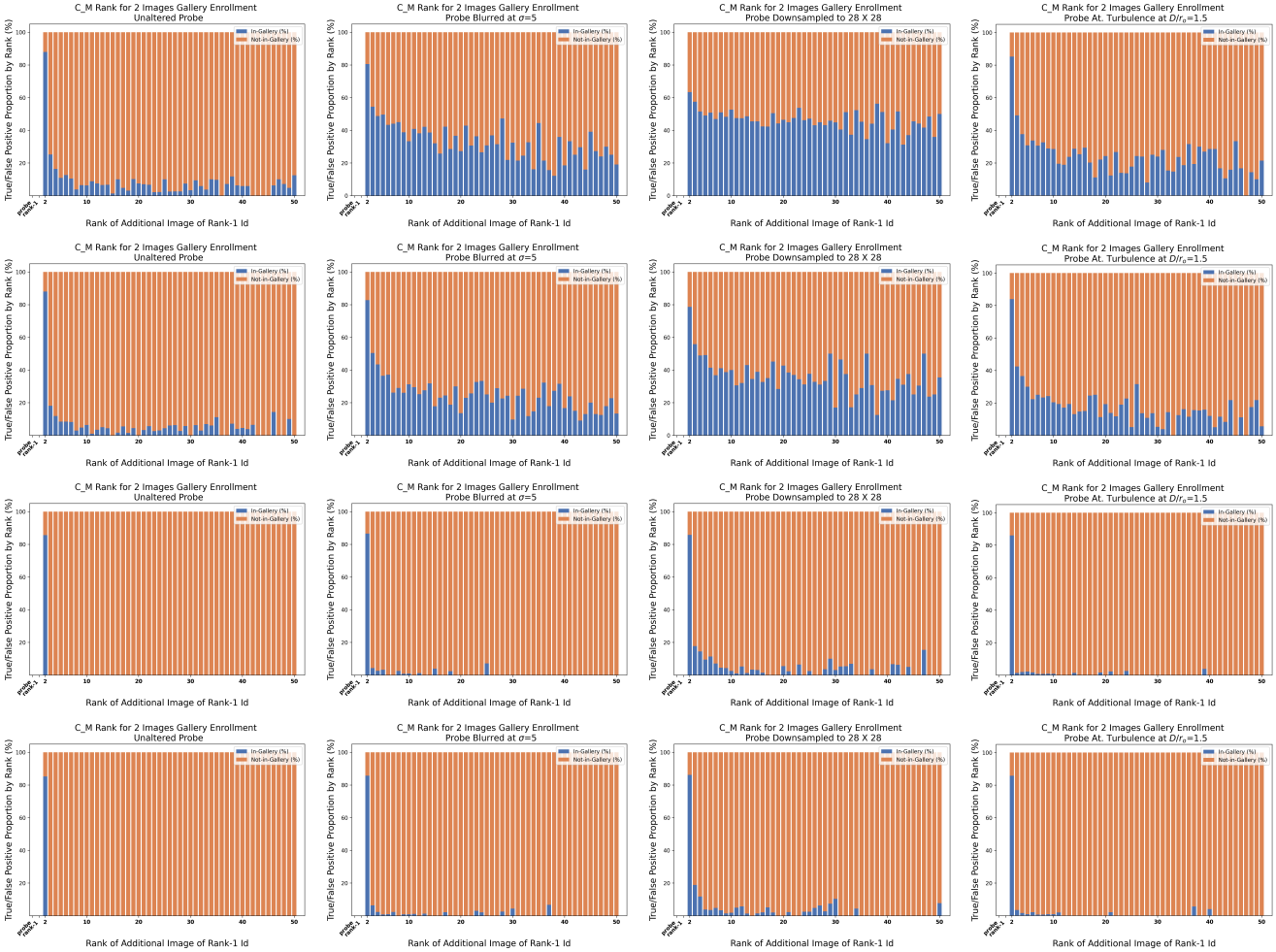
1.1. Caucasian Females



(a) Original Probe (b) Blurred Probe (c) Downsampled Probe (d) Distorted Probe

Figure 1. FaceNet top, ArcFace Second row, TransFace Third row, and AdaFace Last Row

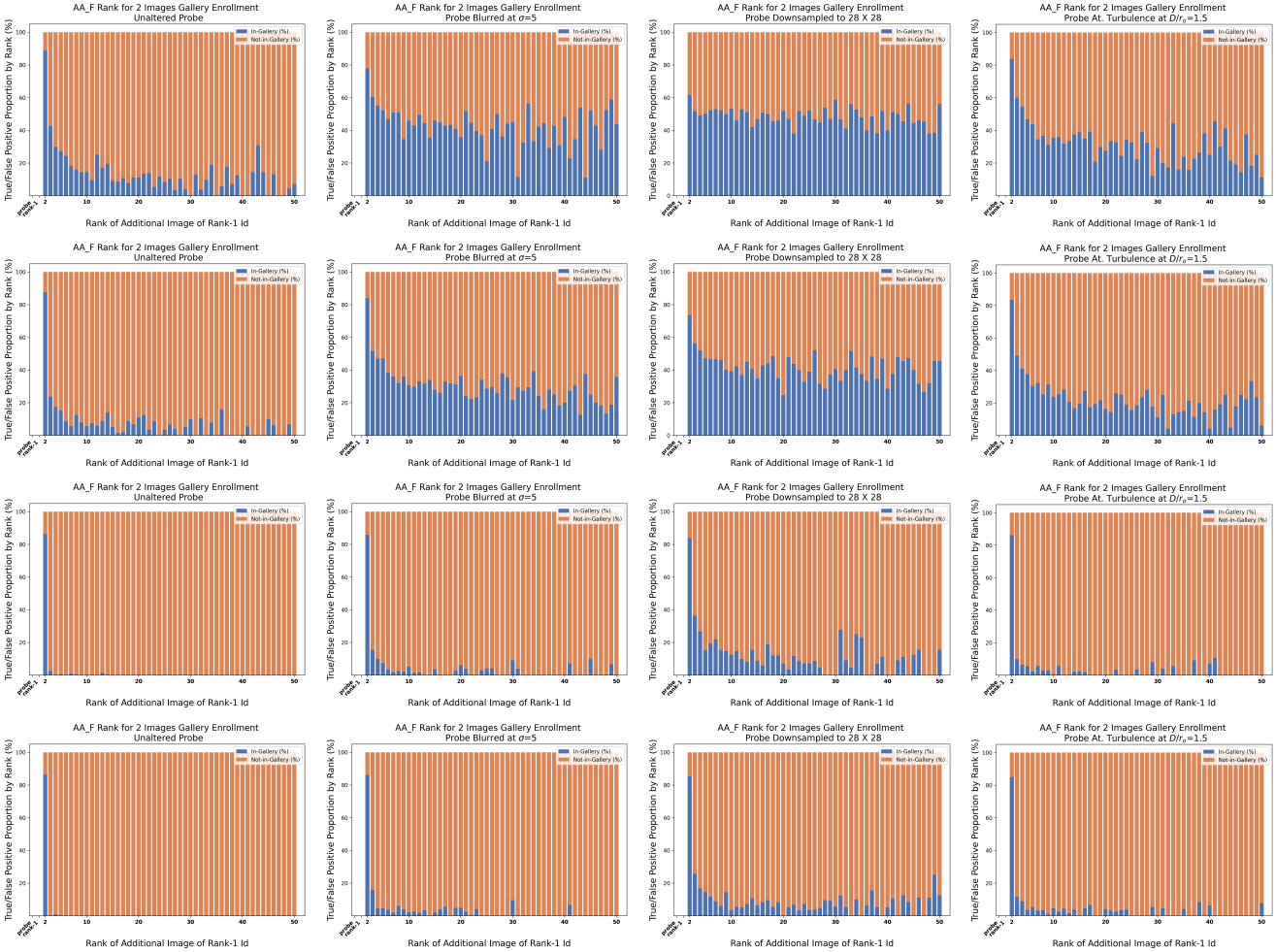
1.2. Caucasian Males



(a) Original Probe (b) Blurred Probe (c) Downsampled Probe (d) Distorted Probe

Figure 2. FaceNet top, ArcFace Second row, TransFace Third row, and AdaFace Last Row

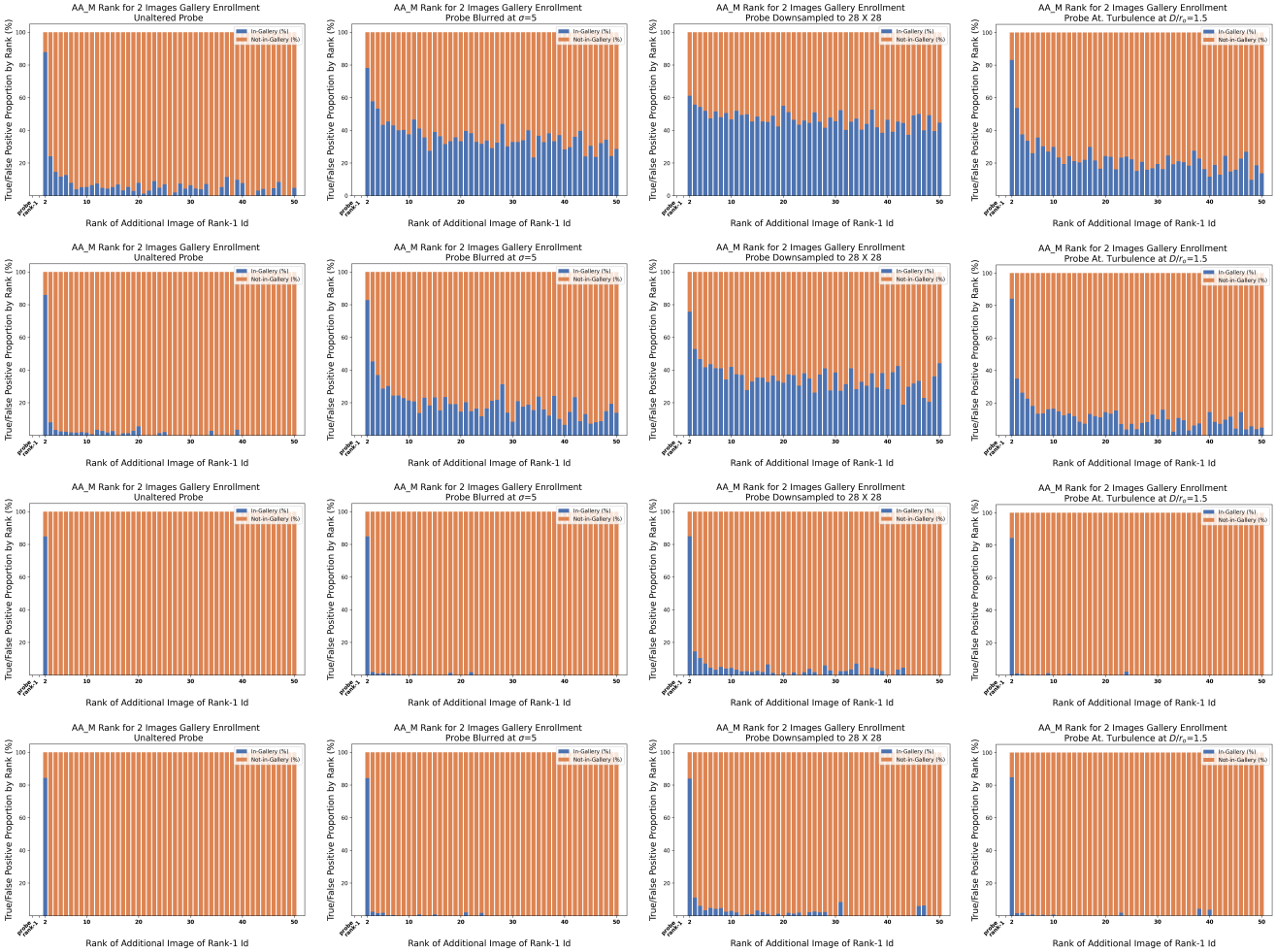
1.3. African-American Females



(a) Original Probe (b) Blurred Probe (c) Downsampled Probe (d) Distorted Probe

Figure 3. FaceNet top, ArcFace Second row, TransFace Third row, and AdaFace Last Row

1.4. African-American Males



(a) Original Probe (b) Blurred Probe (c) Downsampled Probe (d) Distorted Probe

Figure 4. FaceNet top, ArcFace ....

2. Correlation Analysis

In this section, we conduct a feature-to-feature correlation analysis on the rank feature vector both before and after applying rank feature augmentation. As illustrated in Figure 5, the absence of rank feature augmentation results in a high correlation among rank features, a consequence of the ordinal nature of the rank input. This high correlation may lead the model to learn a simplistic decision boundary, which, while potentially effective in controlled settings, lacks robustness in real-world deployment scenarios. To address this, we employ rank feature augmentation, which disrupts the feature-to-feature correlation, as evidenced in Figure 5. By breaking this invariance, the model is compelled to learn a more complex and meaningful non-linear decision boundary, better suited for the challenges of real-world 1-to-many search applications.

(a) Original Probe (b) Blurred Probe (c) Downsampled Probe (d) At. Turbulence

(e) Original Probe (f) Blurred Probe (g) Downsampled Probe (h) At. Turbulence

Figure 5. Correlation analysis of feature rankings before and after row shuffling augmentation.

3. Extended Results on Cardinality Analysis

In the main paper, we presented a cardinality analysis for ArcFace, demonstrating that a rank vector of dimension three—comprising the rank-one match plus three additional images of the rank-one identity—suffices to achieve robust In-gallery vs. Out-of-gallery predictions. Beyond this dimension, performance plateaus, indicating that additional images do not significantly improve classification. In this section, we extend the analysis to FaceNet, AdaFace, and TransFace, showing that these findings generalize consistently across all considered embedding networks. As discussed in the main paper, we focus on African-American males, as this group has the largest dataset, enabling reliable model training even with higher numbers of enrolled images (e.g., six or more per identity).

FaceNet					AdaFace					TransFace				
Num Images per ID	$R^{d_{in}}$	Original	Blurred	Downsampled	Num Images per ID	$R^{d_{in}}$	Original	Blurred	Downsampled	Num Images per ID	$R^{d_{in}}$	Original	Blurred	Downsampled
>6	5	94.65	69.14	54.88	>6	5	98.83	98.08	97.00	>6	5	99.08	98.83	95.66
>5	4	94.28	69.48	53.47	>5	4	98.30	97.96	96.46	>5	4	98.50	97.96	94.96
>4	3	93.51	69.97	54.07	>4	3	98.07	97.47	95.65	>4	3	97.74	97.43	93.23
>3	2	92.20	68.97	50.00	>3	2	96.61	96.08	94.01	>3	2	96.14	95.59	91.95
>2	1	89.75	66.49	50.00	>2	1	90.94	90.45	89.66	>2	1	91.72	90.22	88.57

Table 1. Test of Cardinality using FaceNet, AdaFace, and TransFace. Group: African-American Male.

4. Results of General FR dataset

Our system is primarily designed for galleries of high-quality images, such as state driver's-license or passport-photo collections used by law-enforcement agencies. Nevertheless, assessing performance when gallery image quality varies is also important. Accordingly, we evaluate on LFW and CFP-FP. For LFW, we select identities with at least five images: the most recent image serves as the probe, and the remaining images form the gallery. For CFP-FP, to mimic real-world surveillance scenarios, we identify identities with at least five images that include at least one non-frontal view; from each, we randomly choose one non-frontal/profile image as the probe, and the remaining frontal images constitute the gallery. All probe–gallery sampling follows the methodology in Section 5 of the main paper. Table 2 summarizes the results for LFW and CFP-FP, showing that the In-Gallery/Out-of-Gallery classifier performs effectively with galleries of varying quality and handles non-frontal probe images well.

Dataset	Threshold Classifier	Mean Classifier	Median Classifier	CAFace	Ours
LFW	94.20	71.18	80.59	93.02	**97.06**
CFP-FP	85.30	70.13	79.41	88.40	**88.66**

Table 2. Table 1 [Key: **Best**]