Rank-1 Identity Consistency: Gallery Membership Classification for Operational 1:N Contexts

Anonymous CVPR submission

Paper ID *****

Abstract

In real-world deployments, 1:N face identification systems face a fundamental question: is a probe image's identity enrolled in the gallery or not? We propose the first approach to using consistency of rank-1 identity across multiple matchers as a method to classify the result of 1:N search as in-gallery (ING) or out-of-gallery (OOG). Our "1-consistency" method classifies a probe as ING if all matchers return the same rank-1 identity, and OOG otherwise. We compare its performance to two threshold-based methods: score-thresholding (using raw similarity scores) and gap-thresholding (using the score difference between rank-1 and rank-2 identities).

We evaluate these methods across 12 experimental configurations that systematically vary image quality, gallery enrollment structure, and demographic composition. On average, 1-consistency achieves the highest ING recall (92.0% vs. 72.1% score, 79.6% gap), highest overall accuracy (92.8% vs. 81.4% score, 89.6% gap), and lowest demographic disparities (4.6 pp vs. 10.5 pp score, 8.5 pp gap). Under degraded-probe conditions—the most operationally relevant scenario—1-consistency achieves win margins averaging 16.1 pp for ING recall and 6.1 pp for overall accuracy. With this combination of quality robustness and demographic fairness, 1-consistency is well-suited for real-world contexts like law enforcement investigations—and has the potential to reduce wrongful arrests and enable better allocation of investigative efforts.

1. Introduction

In operational contexts, 1:N face identification systems contend with two major challenges. The first is image quality. Probes frequently originate from surveillance footage and exhibit poor resolution, blur, and other low-quality conditions. Galleries have traditionally been controlled-capture datasets (e.g., databases of state arrest mugshots or driver's licenses), but are increasingly web-scraped. For example,

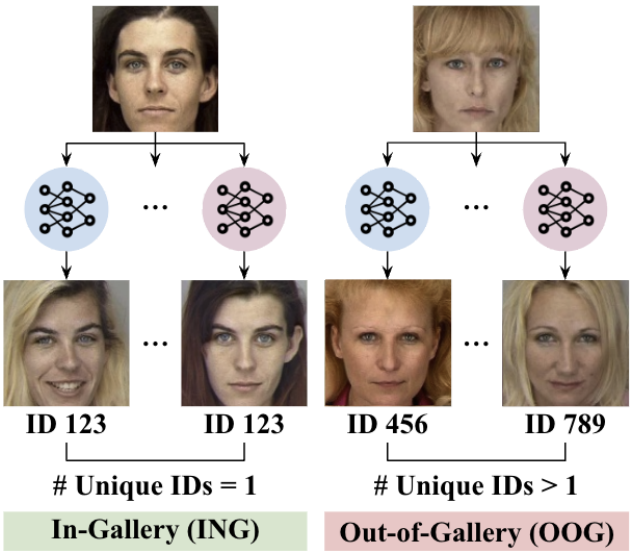


Figure 1. For each probe, multiple matchers perform a 1:N search and return a rank-1 identity. If all rank-1 identities are the same, the probe is labeled in-gallery (ING), else, out-of-gallery (OOG).

Clearview AI's web-scraped database of 70+ billion face images has been used in over 2 million law enforcement searches [1, 8]. Since these images are sourced from "news media, mugshot websites, public social media, and many other open sources," their quality is unknown [1]. Quality has a drastic effect on identification performance: state-of-the-art systems with near-perfect accuracy on high-quality images exhibit not only higher FPIRs (upwards of 20%) but also substantial demographic disparities when probe and/or gallery quality is poor [4, 20, 21].

The second challenge is uncertainty about gallery membership. An investigator searching surveillance footage cannot know in advance whether a subject of interest has ever been photographed and enrolled. This open-set identification problem—determining if a probe's true identity is present in the gallery—is fundamental to investigative workflows. If a probe is out-of-gallery (OOG), any re-

turned candidate is by definition a false positive, which can lead to time being wasted investigating the wrong person and, in the worst case, a wrongful arrest. Conversely, if the probe is in-gallery (ING) but misclassified as OOG, valid investigative leads may be missed. Under ideal conditions (high-quality imagery with guaranteed probe enrollment), modern matchers achieve over 99.9% accuracy even in galleries containing 12–26 million images of 12 million identities [12]. However, distinguishing ING from OOG probes under realistic operational conditions—where quality is degraded and gallery enrollment is unbalanced—remains challenging yet essential.

Existing approaches to gallery membership classification often rely on similarity scores to determine whether to accept or reject returned candidates, e.g., by applying thresholds [13, 15] or estimating probabilities [22–24]. However, since score distributions vary substantially across matcher architectures, quality conditions, and gallery compositions, and exhibit significant demographic differentials, a single global threshold may produce disparate outcomes. That is, a threshold calibrated on one system or dataset may perform poorly if gallery demographics shift, image quality degrades, or the matcher is retrained.

1.1. Contribution of this Work

In this work, we investigate a multi-matcher fusion strategy based on rank stability. While each independently trained matcher instance operates in a distinct embedding space—and may therefore produce varying similarity scores or candidate orderings—consensus on the top-ranked identity signals high certainty. The proposed metric, 1-consistency, classifies a probe based on whether all matcher instances return the same rank-1 identity. Our hypothesis is that ING probes should produce stable rank-1 predictions, while OOG probes will exhibit greater variability as different matcher instances select different gallery impostors.

To our knowledge, this is the first work to use cross-matcher identity consensus for ING/OOG prediction in operational 1:N identification. It is also one of few works to evaluate ING/OOG prediction under controlled variations in image quality and gallery composition, and to report demographic-specific performance differentials. Across the tested conditions, 1-consistency offers a better balance of in-gallery retention and out-of-gallery rejection *and* more consistent cross-demographic performance than threshold-based methods—properties critical for high-stakes investigative deployments.

2. Related Work

Prior work on distinguishing in-gallery from out-of-gallery probes falls into several categories.

Score-thresholding approaches apply fixed or adaptive thresholds on similarity scores to accept or reject candidates [13, 15]. These methods are conceptually simple but highly sensitive to matcher architecture, quality distributions, and gallery composition.

Statistical modeling and Extreme Value Theory (EVT) methods model tail distributions of match scores to estimate the probability that a score originates from a known versus unknown identity [22–24]. Approaches such as W-SVM, Compact Abating Probability (CAP), and Extreme Value Machine (EVM) produce calibrated decision boundaries but require distributional assumptions that may not hold across operational conditions.

Deep open-set recognition methods modify neural network architectures or loss functions to enable rejection of unknown identities [2, 3]. These methods calibrate final-layer activations using EVT but remain dependent on score-based thresholding that is sensitive to quality variation.

Feature-fusion and multi-image aggregation methods stabilize matching by aggregating scores or features across multiple enrolled images [7, 16, 25]. Quality-aware score fusion improves robustness under pose or illumination variation but remains fundamentally limited by reliance on similarity score distributions.

Rank-pattern learning approaches train classifiers to distinguish in-gallery from out-of-gallery probes using rank distributions [5]. This approach shows promise under degraded conditions but is only applicable to identities with more than one image enrolled in the gallery.

In contrast to the approaches described above, we propose a rank-level fusion method that operates on identity consensus across multiple independently trained matcher instances. By leveraging rank stability rather than score calibration, our approach demonstrates improved robustness to quality degradation and more consistent performance across demographic groups.

3. Methodology

In this section, we describe our experimental methodology, including the datasets and degradations used, the matcher configurations, the per-probe processing pipeline, and the classification and evaluation framework. Together, these components allow us to evaluate the robustness of 1-consistency across a range of gallery configurations and quality conditions.

3.1. Datasets

All experiments use images from the MORPH dataset, which consists of mugshot images captured under controlled conditions (i.e., nominally frontal pose, neutral expression, and consistent lighting), along with annotations for race, gender, and age. The images vary in native resolution, ranging from 116×154 to 507×631 pixels, with a modal resolution of 400×480. Each image is close-cropped

around the face region using dlib [17], then aligned and resized to 112×112 pixels to meet CNN input requirements.

We focus on four demographic cohorts within MORPH: Black Female (BLF), Black Male (BLM), White Female (WHF), and White Male (WHM). From each cohort, we select 3,400 identities, forming a "base identity set" of 13,600 total identities used throughout the experiments. For each identity, the most recent image serves as the probe.

All remaining (non-probe) images for each identity constitute the pool from which gallery sets are constructed. Across all MORPH-based experiments, every gallery is either an equal set or a strict superset of the probe identities. Thus, any MORPH probe's identity is always present in the gallery it is matched against; what differs across experiments is the demographic composition of the gallery and the degree of control over per-identity enrollment.

3.2. Degradations

To study quality robustness under controlled and demographically consistent settings, we apply synthetic degradations to MORPH images at both the probe and gallery levels in the experiments where quality is varied. Two degradation types are used, each corresponding to a common problem in probe images taken from surveillance video. First, Gaussian blur is applied to simulate defocus; we choose $\sigma = 4$ as a mid-range severity within the standard NIST 1–7 blur scale [14]. Second, downsampling–upsampling simulates native low-resolution capture: an image is reduced to 18×18 pixels and then resized back to 112×112.

Fig. 2 illustrates the relative quality of the original versus degraded images. Previous work [21] has shown that $\sigma = 4$ blur and 18×18 resolution are equivalent in terms of impact on similarity scores (when performing 1:1 self-matching of each degraded version to the original using ArcFace and AdaFace, the matchers used in this work) and on face image quality assessment scores using MagFace [18], a top-performing metric in NIST evaluations [11].

Table 1. Gallery composition for Experiments 3, 4, and 5.

Exp	Dem	# IDs	# Imgs	Mean	Med
3	BLF	3,400	14,664	4.3	3
	BLM	3,400	19,045	5.6	4
	WHF	3,400	10,228	3.0	2
	WHM	3,400	14,173	4.2	3
4	BLM	21,106	155,715	7.4	5
5	BLF	710	2,909	4.1	3
	BLM	9,414	69,660	7.4	5
	WHF	665	1,891	2.8	2
	WHM	8,846	35,306	4.0	3
	HSM	6,563	24,730	3.8	3
	HSF	494	1,467	3.0	3

3.3. Matchers

We conducted preliminary experiments with both AdaFace and ArcFace. However, ArcFace exhibited substantially worse performance in the quality-degraded scenarios that are central to this work. Given that our focus is on quality-robust gallery membership prediction, we present results only for AdaFace.

We use 10 independently trained instances of AdaFace, each denoted as a *run*. While the runs share the same base architecture (a ResNet-100 backbone with adaptive margin loss trained on Glint360K [10]), their learned embedding spaces are distinct. Consequently, the templates they produce are not interchangeable. For example, performing matching with run X on templates generated by run Y would yield authentic (mated) similarity scores resembling those typically observed for impostor (non-mated) pairs, as demonstrated in [6].

3.4. Per-Probe Processing

In all experiments, each MORPH probe is enrolled in the gallery it is matched against, producing both mated and non-mated candidates for that probe.

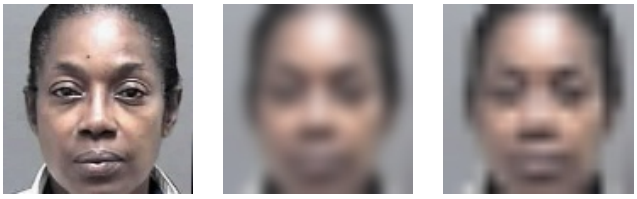
To obtain a probe's in-gallery (ING) outcome, we use the full candidate list (mated + non-mated). To simulate an out-of-gallery (OOG) scenario, we simply discard the mated results and retain only the non-mated candidates—effectively treating the probe as if it had never been enrolled in the gallery. This provides a controlled setup in which ING and OOG cases can be compared directly while holding all other factors constant.

3.4.1. Mean Rank-1 Score (Score-Thresholding)

For each probe in a given configuration, the rank-1 result consists of a single image and its similarity score—the "most similar" image to the probe for that run. Across the 10 independently trained runs, each probe therefore has 10 rank-1 scores. We average these to obtain a single Mean Rank-1 Score per probe per configuration.

3.4.2. k-Consistency (1-Consistency)

From each rank-1 result we also extract the identity of the corresponding image, which we refer to as the *rank-1 ID*—the "most similar identity" to the probe for that



Original | $\sigma = 4.0$ Blur | 18×18 Resolution

Figure 2. Example MORPH image and degraded variants.

run. Each probe thus has 10 rank-1 IDs across the 10 runs. We count the number of unique identities appearing among these 10 values and denote this count as k. A probe with k unique rank-1 IDs is said to be *k-consistent*, where $1 \leq k \leq 10$.

3.4.3. Mean ID Gap (Gap-Thresholding)

We also measure how strongly the rank-1 identity is preferred over all other identities in the gallery. To do this, we identify the "second most similar identity"—the identity (different from the rank-1 ID) that yields the highest similarity score for that run.

This requires special handling when a gallery contains multiple images of the same identity. In such cases, several top-ranked images may all belong to the rank-1 identity. For example, the rank-2, rank-3, ..., rank-X results may all correspond to the same identity as the rank-1 result. Only at rank $X + 1$ do we encounter an image from a different identity. We refer to that identity as the *rank-2 ID*.

Our goal is to quantify the score "distance" between the rank-1 identity and the strongest competing identity. Several definitions are possible (e.g., comparing averages across all images of each identity, or comparing extremal image pairs). For simplicity and consistency, we use the following process: (1) take the best (highest) score among all images belonging to the rank-1 identity (i.e., the rank-1 score), (2) take the best score among all images belonging to the rank-2 identity, (3) subtract the latter from the former. We record this value as the ID Gap for that run. Each probe produces 10 such gap values (one per run), which we average to obtain a single Mean ID Gap per probe per configuration.

3.5. Classification and Evaluation

We perform binary classification of ING/OOG. In each experimental configuration, we calculate the three values—k-consistency, mean rank-1 score, and mean ID gap—for each probe. We determine values of k, mean rank-1 score, and mean ID gap that best separate probes in the baseline conditions of Experiment 1. These thresholds are then used across all other configurations, and our classification rule is as follows: if a probe has a below-threshold value, it is classified as OOG; if above, it is classified as ING.

Because this task involves two classes that are perfectly balanced within each configuration, performance is most naturally expressed in terms of recall for each class:

$$R(\text{ING}) = \frac{\text{\# ING probes correctly labeled ING}}{\text{\# ING probes}} \quad (1)$$

$$R(\text{OOG}) = \frac{\text{\# OOG probes correctly labeled OOG}}{\text{\# OOG probes}} \quad (2)$$

In a balanced binary setting, these recall values share identical denominators and are therefore directly comparable. Because the task is binary and the ING and OOG sets are perfectly balanced, recall alone fully characterizes classification performance: the two recall values share identical denominators, directly reflect the misclassification rates, and determine overall accuracy. Precision, while definable, does not provide additional insight in this setting, since all relevant error patterns are already captured by recall.

Misclassification rates follow immediately from the recalls. The fraction of ING probes misclassified as OOG is $1 - R(\text{ING})$, and the fraction of OOG probes misclassified as ING is $1 - R(\text{OOG})$. Because the classes are balanced, overall accuracy for any method is simply the average of the two recalls.

4. Results and Analysis

To evaluate ING/OOG classification robustness under increasingly realistic operational conditions, we design five experiments that systematically vary quality degradation, cross-demographic matching, or gallery enrollment size.

In this section, we describe the setup and results for each experiment individually. For each, we specify the experimental conditions, provide the gallery configuration details, and analyze classification performance.

4.1. Experiment 1: Baseline

We establish classification thresholds in ideal conditions. Each probe has two mated gallery instances, and is only matched against same-demographic images. Probe and gallery images are original mugshot-quality.

Fig. 3 shows the k-consistency, mean rank-1 score, and mean ID gap distributions for ING and OOG probes. For each method, we determine the threshold that best separates the pair of distributions. Thresholds are indicated by solid blue lines: $k = 1$ for 1-consistency, mean score $= 0.454$ for score-thresholding, and mean gap $= 0.102$ for gap-thresholding. For the score and gap plots, the means of the ING/OOG distributions are shown in solid green/red.

These thresholds are applied to all subsequent experiments without modification. For subsequent experiments, we quantify distribution shifts and classification performance using tabular results rather than distribution plots to conserve space. Tab. 2 provides per-experiment performance metrics (ING recall, OOG recall, and overall accuracy) for each method.

Results. Under baseline conditions, all three methods achieve near-perfect performance, as expected given that thresholds were selected to optimally separate ING and OOG in this setting. ING recall exceeds 99% for all three methods, with 1-consistency marginally best at 99.8%. OOG recall is 99.1% for both thresholding methods, while 1-consistency is lower at 91.8%. Overall accuracy reflects this pattern: threshold-based methods achieve 99.2–99.4%,

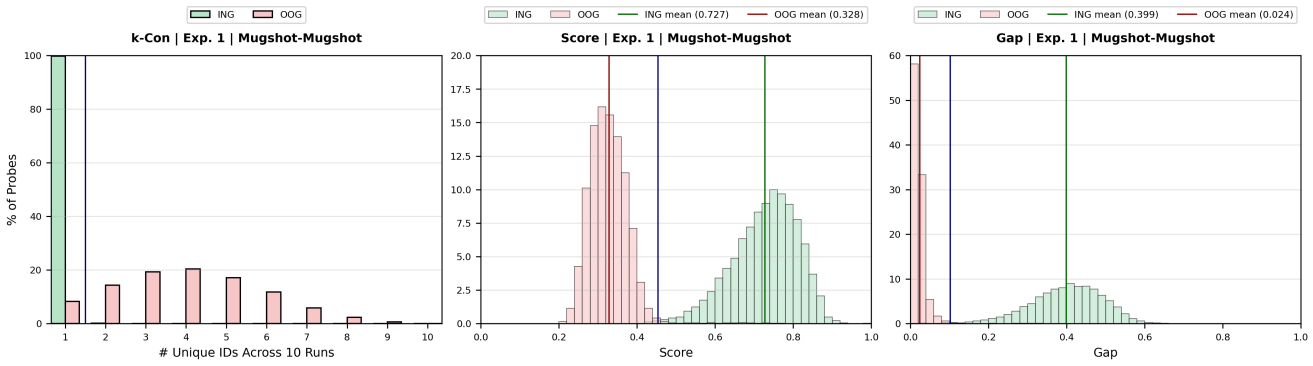


Figure 3. Baseline comparison for good-quality images and single-demographic galleries with balanced enrollment.

Table 2. Average performance metrics for the three classification methods. Best method(s) per configuration and metric highlighted in bold green. Blur (BL) and resolution (RE) degradations indicated where applicable.

	Gal Composition		Degradation		ING Recall			OOG Recall			Accuracy		
Exp	**N Imgs**	**N Dems**	**Probe**	**Gal**	**1-Con**	**Score**	**Gap**	**1-Con**	**Score**	**Gap**	**1-Con**	**Score**	**Gap**
1	2	1	—	—	**99.8**	99.7	99.4	91.8	**99.1**	**99.1**	95.8	**99.4**	99.2
2	2	1	BL	—	**87.9**	44.7	71.4	95.8	**99.9**	99.6	**91.9**	72.3	85.5
			RE	—	**87.2**	47.1	70.6	95.6	**99.9**	99.6	**91.4**	73.5	85.1
			—	BL	**92.1**	45.7	78.6	95.9	**99.9**	99.7	**94.0**	72.8	89.2
			—	RE	**91.6**	49.7	77.2	95.8	**99.9**	99.7	**93.7**	74.8	88.5
			BL	BL	83.6	**97.0**	71.0	90.0	31.0	**99.3**	**86.8**	64.0	85.2
			RE	RE	83.8	**91.2**	70.8	91.1	63.6	**99.4**	**87.5**	77.4	85.1
3	2-17	4	—	—	**99.9**	**99.9**	99.5	92.1	98.8	**99.1**	96.0	**99.4**	99.3
			RE	—	**89.3**	55.9	73.9	96.1	**99.8**	99.7	**92.7**	77.9	86.8
4	1-108	1	—	—	**100**	**100**	99.9	94.2	96.1	**99.6**	97.1	98.1	**99.8**
			RE	—	**87.9**	71.9	67.2	96.9	99.7	**99.9**	**92.4**	85.8	83.6
5	1-108	6	RE	—	**91.3**	62.3	75.5	97.2	**100**	99.9	**94.3**	81.2	87.7
Averages Across All Configurations					**92.0**	72.1	79.6	94.4	90.7	**99.6**	**92.8**	81.4	89.6

while 1-consistency achieves 95.8%. Demographic differences in accuracy are minor across all three methods (1–2 percentage points).

These results demonstrate that under ideal quality and gallery enrollment conditions, distinguishing ING from OOG probes is straightforward using rank-1 scoring information alone, whether raw similarity scores or score gaps relative to competing identities. However, such ideal conditions rarely hold in operational deployments. In the following experiments, we evaluate classification performance under increasingly realistic and challenging scenarios: degraded image quality, varied gallery structures, and cross-demographic matching.

4.2. Experiment 2: Variable Image Quality

To what extent does quality degradation—in probe images, gallery images, or both—impact classification performance? As in experiment 1, the gallery is single-demographic with two enrolled images per probe.

We evaluate three quality scenarios: (1) probe-only degradation, representing the most operationally relevant scenario where surveillance probes are matched against controlled-capture mugshot databases, (2) gallery-only degradation, an operationally unlikely scenario included here for completeness, and (3) symmetric degradation, which may arise from the use of uncontrolled web-scraped galleries. For each scenario, we apply both blur (BL: $\sigma = 4$) and resolution (RE: 18×18) degradations. This design isolates the individual and joint effects of probe and gallery quality on classification accuracy.

Results. Quality degradation substantially impacts classification performance, with different methods showing distinct failure modes. For probe-only degradation, 1-consistency maintains strong ING recall (87.2–87.9%). Score-thresholding collapses catastrophically (44.7–47.1%), while gap-thresholding performs moderately

Table 3. Best and worst demographic by accuracy for each method. Diff = percentage point difference.

Exp	Degradation		1-Consistency			Score-Threshold			Gap-Threshold		
	Probe	Gal	Best	Worst	Diff	Best	Worst	Diff	Best	Worst	Diff
1	—	—	BLF	WHF	2.1	BLM	WHF	1.0	BLF	WHF	1.6
2	RE	—	BLM	BLF	7.6	BLM	WHF	9.9	BLM	BLF	12.2
	—	RE	BLM	WHF	2.4	BLM	WHF	12.4	BLM	BLF	12.0
	RE	RE	BLM	BLF	6.1	WHM	BLF	24.2	BLM	BLF	9.5
3	—	—	WHM	WHF	1.3	WHM	WHF	1.1	WHM	WHF	1.9
	RE	—	BLM	BLF	8.0	BLM	WHF	14.3	BLM	BLF	13.5

(70.6–71.4%), still substantially lower than 1-consistency. OOG recall remains high for all methods (95.6–99.9%). Overall accuracy favors 1-consistency (91.4–91.9%) over both score-thresholding (72.3–73.5%) and gap-thresholding (85.1–85.5%). Demographic disparities increase substantially for all methods but remain lowest for 1-consistency (7.6–8.0 pp) compared to score-thresholding (9.9–12.5 pp) and gap-thresholding (12.9–13.5 pp).

For gallery-only degradation, the pattern is similar: 1-consistency achieves ING recall of 91.6–92.1% while threshold methods again struggle (45.7–49.7% for score, 77.2–78.6% for gap), demonstrating that both probe and gallery quality independently affect threshold-based method performance.

For symmetric degradation (both probe and gallery degraded), score-thresholding achieves the highest ING recall (91.2–97.0%) but catastrophically fails on OOG recall (31.0–63.6%), resulting in poor overall accuracy (64.0–77.4%). Gap-thresholding achieves near-perfect OOG recall (99.3–99.4%) but lower ING recall (70.8–71.0%). 1-consistency provides the most balanced performance, achieving the highest overall accuracy (86.8–87.5%) and lowest demographic disparities (6.2–7.4 pp) despite not winning either individual recall metric.

These results show that threshold-based methods are highly sensitive to quality degradation, with score-thresholding primarily failing on ING recall and occasionally on OOG recall under dual degradation. In contrast, 1-consistency maintains more balanced performance across quality conditions, trading modest OOG recall for substantially higher ING recall and lower demographic disparities.

Given that probe-only degradation is the most operationally relevant scenario and that blur and resolution degradations yield comparable results within each method for this case, subsequent experiments use resolution-degraded probes.

4.3. Experiment 3: Variable Gallery Composition

To what extent does gallery composition—variable enrollment size and demographic heterogeneity—impact classification performance? In this experiment, the gallery grows in size: it comprises 58,110 total images of all four demographics. Each identity has 2–17 enrolled images, and the average count differs across demographics: 3 for WHF, 4 for WHM and BLF, and 6 for BLM. Note that the number of represented *identities* is still balanced across demographics.

We consider the impact of gallery composition for original mugshot-quality probes and resolution-degraded probes. For original probes, we compare results to those of experiment 1. For degraded probes, we compare results to those of experiment 2.

Results. When probes are mugshot-quality, gallery composition has negligible impact: across all method/metric pairings, the average difference from Experiment 1 baseline is only 0.1 percentage points. All three methods achieve near-perfect performance (96.0–99.4% accuracy).

When probes are resolution-degraded, gallery composition introduces modest variation: the average difference from Experiment 2 is 2.5 percentage points. The largest shifts occurs for ING recall, which *improves* for all methods: up 2.1 pp for 1-consistency, 3.3 pp for gap, and 8.8 for score (though the latter two still remain low overall). 1-consistency still maintains the highest ING recall (89.3%) and overall accuracy (92.7%), with the lowest demographic disparities (8.0 pp vs. 13.7–14.3 pp for threshold methods).

These results confirm that, thus far, image quality remains the dominant factor determining classification performance.

4.4. Experiment 4: Larger-Scale Gallery

This experiment evaluates classification under a large-scale, uncontrolled gallery structure: 1–108 images per identity, single-demographic composition (BLM only), and probe-only degradation. The gallery includes 21,106 BLM identities—the 3,400 base identities (each contributing all remaining non-probe images, typically 2–17) plus 17,706 additional BLM identities drawn from the full MORPH dataset—yielding 155,715 total images. This configuration maximizes both gallery size and within-identity variability while maintaining demographic homogeneity, stress-testing classification methods under large-scale conditions repre-

sentative of operational databases.

Results. Under mugshot-quality conditions, all three methods achieve near-perfect performance, comparable to Experiment 1 baseline (controlled single-demographic gallery with 2 images per identity). Comparing Experiment 4 to Experiment 1: 1-consistency ING recall remains essentially identical (100.0% vs. 99.8%), score-thresholding ING recall increases marginally (100.0% vs. 99.7%), and gap-thresholding ING recall increases slightly (99.9% vs. 99.4%). Overall accuracy ranges from 97.1–99.8% in Experiment 4 versus 95.8–99.4% in Experiment 1. The large gallery size (155,715 images, 21,106 identities) does not impair classification when image quality is high.

Under degraded probe conditions, we compare to Experiment 2 (controlled single-demographic gallery, resolution-degraded probes). 1-consistency ING recall remains nearly stable (87.9% in Experiment 4 vs. 87.2% in Experiment 2), demonstrating robustness to gallery scale. Score-thresholding ING recall improves substantially from 47.1% to 71.9%—a 24.8 pp increase. Gap-thresholding ING recall decreases slightly from 70.6% to 67.2%—a 3.4 pp decline. Notably, this represents the first configuration where gap-thresholding performs worse than score-thresholding on ING recall. OOG recall remains high for all methods (96.9–99.9%). Overall accuracy favors 1-consistency (92.4%) over score-thresholding (85.8%) and gap-thresholding (83.6%).

These results demonstrate that 1-consistency's rank-stability approach remains largely insensitive to gallery scale and within-identity variability, while threshold-based methods show moderate sensitivity—score-thresholding improving and gap-thresholding declining under large-scale conditions.

4.5. Experiment 5: Simulated Operational Gallery

This experiment simulates an operational investigative scenario with the most realistic conditions tested: uncontrolled gallery structure (1–108 images per identity), cross-demographic composition reflecting actual operational distributions, and both synthetically degraded and real surveillance-quality probes. The gallery includes two additional demographics, Hispanic females and males (HSF, HSM) and 13,092 additional identities, for a total of 136,074 total images. The demographic composition approximates the distribution observed in U.S. state and federal prisons [19]: 35% BLM, 33% WHM, 25% HSM, 3% BLF, 3% WHF, 2% HSF. This configuration represents a realistic mugshot database used for law enforcement deployments.

We evaluate two probe types: (1) resolution-degraded MORPH probes (1,500 WHM and 1,500 BLM), and (2) real surveillance imagery from the QMUL SurvFace dataset [9]. For QMUL, we selected a subset with square aspect ratios and at least 18×18 resolution to enable resizing to 112×112 for matcher ingestion and facilitate comparison with synthetically degraded probes.

Results. For resolution-degraded MORPH probes, we compare to Experiment 3 (semi-controlled multi-demographic gallery, resolution-degraded probes). 1-consistency ING recall improves slightly from 89.3% to 91.3%, while score-thresholding improves from 55.9% to 62.3%, and gap-thresholding improves from 73.9% to 75.5%. OOG recall remains high for all methods, with score-thresholding achieving perfect classification (100.0%). Overall accuracy favors 1-consistency (94.3%) over score-thresholding (81.2%) and gap-thresholding (87.7%).

The modest improvements across all methods suggest that the operational gallery structure—despite its larger scale (136K vs. 58K images), demographic imbalance, and broader demographic coverage (6 vs. 4 groups)—does not substantially impair classification compared to the more controlled multi-demographic setting of Experiment 3. This reinforces that image quality remains the dominant factor.

For QMUL surveillance probes, all three methods achieved 100% OOG recall, correctly rejecting every probe. However, the QMUL images exhibited severe degradation across multiple dimensions—extreme occlusions (hats, glasses), substantial pose variation, and poor lighting—representing quality conditions that would ideally be flagged during human review and excluded from operational searches. This unanimous rejection suggests that when probe quality is sufficiently poor, all methods reliably detect unreliability, though the scenario provides limited discriminative insight into method differences.

These results demonstrate that 1-consistency maintains its advantages under the most operationally realistic conditions tested, achieving the highest accuracy and balanced performance across ING and OOG recall even in large-scale, demographically heterogeneous, structurally uncontrolled galleries.

5. Discussion

Across all experimental configurations, 1-consistency demonstrates superior and more robust performance for ING/OOG classification compared to threshold-based methods. This advantage is most pronounced under the conditions that matter most for operational deployment: degraded probe quality matched against high-quality gallery databases. The results reveal three key findings.

(1) Image quality drives classification performance. Probe degradation causes threshold-based methods to collapse catastrophically on ING recall (often by 40–50 percentage points), while 1-consistency maintains stable performance (typically within 10–15 pp of baseline). For OOG recall, threshold methods maintain marginal advan-

tages (typically 3–5 pp higher than 1-consistency), but this slight gain comes at an unacceptable operational cost: discarding 40–50% of valid in-gallery probes undermines the fundamental purpose of investigative identification systems. For demographic fairness, probe degradation increases disparities across all methods, but 1-consistency maintains the smallest increases (5–6 pp from baseline) compared to score-thresholding (9–11 pp) and gap-thresholding (11–12 pp), consistently achieving the lowest demographic differentials under challenging conditions.

(2) Classification is relatively stable across gallery structures. We evaluated galleries varying in number of images per identity (fixed at 2 versus variable from 1–108), number of identities (from 3,400 to 26,692), and demographic composition (from one to six represented groups). For all three methods, these variations produced only minor performance effects.

However, our experiments do not extend to galleries scaled by orders of magnitude beyond those tested, e.g., 100× or 1000× more identities within the same demographic, as may occur in certain state-level deployments. While our results suggest that quality effects dominate within moderate-to-large gallery settings, performance under extreme-scale single-demographic galleries remains an open question for future study.

(2) 1-consistency achieves the best balance across competing objectives. It wins on ING recall in 10 of 15 configurations (by 16.3 pp on average), on overall accuracy in 12 of 15 configurations (by 6.7 pp on average), and on demographic fairness in 8 of 10 configurations reporting demographics (by 4.2 pp on average). Threshold-based methods lead only on OOG recall—and even there, 1-consistency trails by just 5.1 pp on average. This asymmetry matters operationally: failing to retain valid in-gallery probes (low ING recall) directly undermines investigative workflows, while the modest OOG recall tradeoff represents a far less costly error mode.

(3) The most operationally relevant quality scenario—the comparison of degraded probes to a good-quality gallery—is where 1-consistency excels. Across the six configurations testing this scenario, 1-consistency achieves 15–27 pp higher ING recall than threshold methods, 6–12 pp higher overall accuracy, and 2–6 pp lower demographic disparities. Threshold methods' OOG recall advantage in these scenarios is negligible (0.1–0.3 pp).

This performance gap reflects a fundamental distributional problem, not merely a calibration issue. Distributional separability analysis (Table 7) reveals how degradation affects each method's ability to distinguish ING from OOG probes. Under probe degradation, threshold-based separability *collapses* by 60–70% from baseline, while 1-consistency's rank-stability signal actually becomes *more* discriminative, improving by 12–29%. This means threshold-based methods cannot be improved through recalibration—the underlying score distributions have collapsed to the point where no single threshold can effectively separate ING from OOG probes.

5.1. Operational Implications

The one scenario where threshold-based methods show competitive performance is the idealized baseline: mugshot-quality probes and galleries with controlled enrollment. Under these conditions, all methods perform nearly identically (differences under 1 pp). However, such conditions rarely hold in operational systems, where probe quality degradation is the norm rather than the exception.

By operating on rank agreement across independently trained matcher instances rather than on matcher-specific score distributions, 1-consistency achieves greater invariance to the quality perturbations that characterize real investigative deployments. This combination of higher accuracy, better demographic fairness, and structural robustness to degradation makes 1-consistency the most operationally viable approach for distinguishing in-gallery from out-of-gallery probes in realistic face identification systems.

6. Conclusion

We investigated 1-consistency, a rank-stability approach for distinguishing in-gallery from out-of-gallery probes in face identification systems. Across 15 experimental configurations spanning controlled to uncontrolled galleries, mugshot to degraded image quality, and single- to multi-demographic compositions, 1-consistency consistently outperformed threshold-based methods on the metrics that matter most for operational deployment: in-gallery recall, overall accuracy, and demographic fairness.

The advantage is most pronounced under realistic operational conditions—degraded probe quality against high-quality gallery databases—where threshold-based methods suffer catastrophic performance losses. Distributional analysis reveals this is not a calibration problem but a structural collapse: similarity score distributions lose discriminative power under degradation (60–70% reduction in separability), while rank stability actually becomes more discriminative (12–29% improvement).

Remarkably, even under the most challenging dual-degradation conditions, 83.6–87% of true in-gallery probes achieve perfect rank-1 consensus across all 10 independently trained matcher instances, while 90–91% of out-of-gallery probes fail to achieve consensus. This demonstrates that rank stability provides a fundamentally more robust signal for gallery membership than similarity scores, making 1-consistency the most operationally viable approach for open-set face identification under realistic deployment conditions.

References

[1] Clearview AI. Clearview ai 2.0, 2025. Accessed: 2025-02-20. 1

[2] Abhijit Bendale and Terrance E. Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1893–1902, 2015. 2

[3] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1563–1572, 2016. 2

[4] Aman Bhatta, Gabriella Pangelinan, Michael C King, and Kevin W Bowyer. Impact of blur and resolution on demographic disparities in 1-to-many facial identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 412–420, 2024. 1

[5] Aman Bhatta, Maria Dhakal, Michael C King, and Kevin W Bowyer. Are you in or out (of gallery)? wisdom from the same-identity crowd. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5946–5955, 2025. 2

[6] Aman Bhatta, Michael C King, and Kevin W Bowyer. Deep cnn face matchers inherently support revocable biometric templates. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2025. 3

[7] Navaneeth Bodla, Jingxiao Zheng, Hongyu Xu, Jun-Cheng Chen, Carlos Castillo, and Rama Chellappa. Deep heterogeneous feature fusion for template-based face recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 586–595, 2017. 2

[8] Chris Burt. Clearview facial recognition searches double, database reaches 50b images. *Biometric Update*, 2024. Accessed: 2025-02-20. 1

[9] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Surveillance face recognition challenge. *arXiv preprint arXiv:1804.09691*, 2018. 7

[10] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444, 2021. 3

[11] Federal Office for Information Security (BSI). Open Face Image Quality (OFIQ): Short Report. Technical report, Federal Office for Information Security (BSI), 2021. Accessed: 2025-07-08. 3

[12] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face Recognition Vendor Test (FRVT) Part 2: Identification. https://doi.org/10.6028/NIST.IR.8271, 2019. 2

[13] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test (FVRT): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019. 2

[14] Patrick Grother, Mei Ngan, and Elham Tabassi. Face Analysis Technology Evaluation (FATE) Part 11: Face Image Quality Vector Assessment. https://doi.org/10.6028/NIST.IR.8485, 2023. 3

[15] Manuel Gunther, Steve Cruz, Ethan M. Rudd, and Terrance E. Boult. Toward open-set face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 71–80, 2017. 2

[16] Minchul Kim, Feng Liu, Anil K Jain, and Xiaoming Liu. Cluster and aggregate: Face recognition with large probe set. *Advances in Neural Information Processing Systems*, 35:36054–36066, 2022. 2

[17] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 3

[18] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14234, 2021. 3

[19] Derek Mueller and Rich Kluckow. Prisoners in 2023 – statistical tables. Technical Report NCJ 310197, U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, 2025. 7

[20] Gabriella Pangelinan, Aman Bhatta, Haiyu Wu, Michael C King, and Kevin W Bowyer. Analyzing the impact of demographic and operational variables on 1-to-many face id search. *IEEE Transactions on Technology and Society*, 2024. 1

[21] Gabriella Pangelinan, Michael C. King, and Kevin W. Bowyer. When probe and gallery are low quality: Decreasing accuracy and increasing demographic disparities in 1:n identification. In *Proceedings of the Winter Conference on Applications of Computer Vision*, 2026. To appear. 1, 3

[22] Walter J. Scheirer, Anderson Rocha, Ross J. Micheals, and Terrance E. Boult. Meta-recognition: The theory and practice of recognition score analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1689–1695, 2011. 2

[23] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2012.

[24] Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2317–2324, 2014. 2

[25] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4362–4371, 2017. 2