Pairwise Accuracy Is Not Security: Masterface Attacks Expose a Structural Vulnerability in Face Verification

Technical Report

Ehsan Nazari

Abstract

Face verification systems are often treated as secure when they achieve near-perfect pairwise accuracy. We show that this conclusion does not follow: a model can achieve 99.65% pairwise accuracy while nine masterface *embeddings—adversarial points engineered to exploit dense regions of the embedding space—collectively match 47.2% of all identities in the same evaluation setting, with the best single embedding covering 9.3%. This work argues that pairwise verification accuracy is insufficient as a security metric. Models trained with state-of-the-art objectives—softmax angular margins (ArcFace, CosFace, AdaFace) and triplet losses (FaceNet)—achieve near-perfect pairwise accuracy yet still contain regions where many unrelated identities fall within the same decision neighborhood. We provide three lines of evidence. First, large-scale masterface coverage can coexist with high pairwise accuracy across multiple models. Second, seemingly minor pipeline choices can change the magnitude of this vulnerability by up to two orders of magnitude while leaving pairwise accuracy almost unchanged. Third, even under the most conservative configuration we test, small randomly selected groups of identities still share a common covering point with near-perfect success, indicating a model-level floor that pipeline choices do not remove. Finally, we show that these embeddings can be realized as plausible face images, retaining 88–97% of their embedding-space coverage under the most vulnerable pipeline and enabling practical attacks such as digital identity fraud. We propose that face verification systems report masterface coverage and joint-threshold attack curves alongside standard pairwise accuracy. Code to reproduce all experiments is available at* https://github.com/enazari/masterface-attacks.

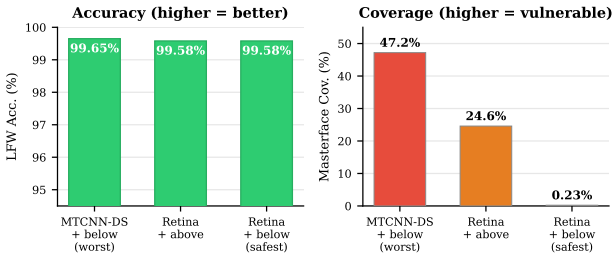

Figure 1. **Same model, three pipelines: accuracy barely moves; masterface coverage swings two orders of magnitude.** CosFace under three pipeline configurations. Pairwise accuracy (left) is uniformly ≈99.6%—the standard benchmark sees no difference. Masterface coverage (right) varies from 47.2% to 0.23%—the attack surface differs by a factor of more than 200.

1. Introduction

Face verification is commonly evaluated as a pairwise decision problem: given two face images, the system decides whether they belong to the same identity. The dominant benchmark protocol reports accuracy on Labeled Faces in the Wild (LFW), a standard face-verification dataset, and often fixes the operating threshold using a target false-acceptance rate (FAR), the fraction of impostor pairs incorrectly accepted as matches [11]. These numbers are important, but they measure average pairwise behavior. They do not directly measure whether a single adversarial face can match many enrolled identities.

This distinction matters because prior work on *masterfaces* showed that face recognition systems can be vulnerable to images that match unusually many identities [10]. Such attacks are not conventional pairwise errors sampled at random; they exploit dense regions of the embedding space, where many unrelated identities lie close enough to a common point to fall inside the same decision threshold. Some configurations appear more robust to these attacks than others, but the field still lacks a clear account of which parts of the vulnerability come from evaluation and deployment

choices, and which parts are intrinsic to the learned embedding models.

We study this gap systematically. We evaluate six face-mapping models—ArcFace, CosFace, AdaFace-IR101, AdaFace-ViT, and two FaceNet variants—under multiple face-alignment and threshold-selection configurations on LFW. We use a two-phase attack. Phase 1 searches directly in the embedding space for points that cover many identities. Phase 2 realizes those embeddings as plausible face images by optimizing pixels toward the target embedding. This separation lets us distinguish the geometric vulnerability of the embedding space from the additional difficulty of producing a visible attack image.

Our central finding is that high pairwise accuracy and masterface security are decoupled. Figure 1 illustrates the disconnect for CosFace: pairwise accuracy remains near 99.6% across three pipeline configurations, while nine masterface embeddings cover anywhere from 0.23% to 47.2% of identities. The same model can therefore look equally accurate under the standard benchmark while exposing radically different attack surfaces.

The results also show that the vulnerability has two layers. The first layer is pipeline sensitivity: alignment and threshold-selection choices can amplify or suppress large-scale masterface coverage by up to two orders of magnitude without a corresponding change in pairwise accuracy. The second layer is a model-level floor: even under the most conservative configuration we test, randomly selected groups of 2–4 identities almost always share a common covering point. We formalize this small-group diagnostic as the Joint-Threshold Attack (JT-Attack), which asks whether one embedding can simultaneously fall within the verification threshold of several chosen identities.

This work makes the following contributions:

- **An optimization-based masterface attack.** We present a two-phase method that first searches the embedding space for high-coverage points using LM-MA-ES and then realizes those points as face images through pixel-space optimization. On FaceNet-CASIA, the method reaches 49.957% image-space coverage, surpassing the 43.82% reported by Shmelkin et al. [10] with a simpler, non-generative pipeline.

- **A pipeline-sensitivity analysis.** We show that threshold-selection direction and face-alignment strategy can each change masterface coverage by up to two orders of magnitude while leaving pairwise accuracy nearly unchanged. These choices are therefore security-critical even though standard pairwise benchmarks barely expose them.

- **Evidence for a model-level vulnerability floor.** We introduce the JT-Attack and show that, even after choosing the most conservative pipeline configuration, every model tested still admits a covering embedding for any 2–4 randomly selected identities with near-perfect success. This suggests that dense-region vulnerability is not only a pipeline artifact; it is also a structural property of current face embedding models.

The remainder of this work is organized as follows. Section 2 defines the verification pipeline and evaluation protocol. Section 3 reviews masterface attacks and related security evaluations. Section 4 describes our optimization-based attack. Section 5 presents the experimental setup and headline coverage results. Section 6 isolates the effects of threshold selection and alignment. Section 7 introduces the JT-Attack and analyzes the model-level floor. Section 8 evaluates image-space realization, and Sections 9–10 discuss implications and conclude.

2. Background

A face verification pipeline consists of three stages: (1) a *face detector* that locates and aligns faces in an image, (2) a *face mapper* or embedding model that projects the aligned face into a high-dimensional embedding space, and (3) a *decision rule* that declares a match if the distance between two embeddings falls below a threshold τ.

Face detectors and alignment. MTCNN [12] is a widely used cascaded CNN detector. RetinaFace [2] uses single-stage detection with 5-point landmark alignment, producing more consistently aligned faces. Because the detector determines the cropped and aligned image passed to the mapper, this choice can change the distribution of embeddings even when the face-mapping model is fixed.

Threshold selection at a target FAR. The decision threshold is typically selected to satisfy a target False Acceptance Rate (FAR), where FAR is the fraction of impostor pairs incorrectly accepted as matches. We focus on the common FAR$\sim$0.001 operating point. For each LFW fold, we sweep 10,000 evenly spaced candidate thresholds; each threshold yields a corresponding FAR, but in practice no threshold achieves FAR $= 0.001$ exactly. One therefore selects the closest threshold from one of two directions: the *from-below* threshold is the closest candidate with FAR < 0.001 and is more conservative, while the *from-above* threshold is the closest candidate with FAR > 0.001 and is more permissive. The final threshold is the average across the ten LFW folds.

Pairwise evaluation and its limits. Standard benchmarks, such as LFW accuracy at a target FAR, evaluate whether randomly sampled genuine and impostor pairs are correctly classified. This measures average-case separation between identities. It does not measure worst-case neighborhood density: whether there exist regions where many unrelated identities cluster close enough for a single adversarial point to match all of them. Our results show that these

two properties are decoupled.

3. Related Work

Face verification has advanced rapidly through deep embedding models trained to separate identities in a metric space. Angular-margin objectives, including the families represented by ArcFace, CosFace, and AdaFace, encourage embeddings of the same identity to cluster tightly while pushing different identities apart. Triplet-loss systems such as FaceNet impose a related metric-learning objective by comparing anchor, positive, and negative samples. These methods achieve very high LFW accuracy and are often evaluated through pairwise verification protocols [11].

Security evaluation requires a different lens. Pairwise benchmarks estimate the probability that a random genuine or impostor pair is classified correctly, whereas an adversary searches for worst-case inputs. Masterface attacks occupy this worst-case setting. Shmelkin et al. [10] introduced the masterface concept and demonstrated that a generated face can match many identities in a face recognition system. Their approach uses a GAN-based generation pipeline and evaluates image-space coverage.

Our work builds on this line of attack but separates two questions that are usually entangled. First, does the embedding space itself contain high-coverage points? Second, can those points be realized as plausible face images? We answer the first question through a direct embedding-space search and the second through pixel-space optimization. This separation lets us analyze pipeline sensitivity and model-level dense-region structure independently of the image generator. The contribution is therefore not only higher raw coverage, although we report a direct comparison in Section 8; it is the evidence that high pairwise accuracy can coexist with a large adversarial attack surface, and that part of this surface remains even under conservative pipeline choices.

4. The Masterface Attack

This section describes our method for generating masterface embeddings and corresponding images.

4.1. Problem Setup and Notation

Let $FM : \mathcal{I} \to \mathbb{R}^d$ denote a face-mapping model that projects a face image $\mathcal{I}$ into a d-dimensional embedding space. Given a collection of face images, we denote by

$$\mathcal{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_N\}, \quad \mathbf{e}_i = FM(\mathcal{I}_i),$$

the corresponding set of face embeddings. For many prominent face mappers, this space is ℓ_2-normalized, forming a hypersphere in typically 512 dimensions. Distances in the embedding space are measured using a metric $d(\cdot, \cdot)$, such as cosine or Euclidean distance. A face verification system declares two embeddings to match if their distance is below a fixed decision threshold τ. For any point $\mathbf{x} \in \mathbb{R}^d$, we define its τ-neighborhood relative to a set of embeddings $\mathcal{E}$ as

$$\mathcal{N}_\tau^{\mathcal{E}}(\mathbf{x}) = \{\mathbf{e}_i \in \mathcal{E} \mid d(\mathbf{x}, \mathbf{e}_i) \leq \tau\}.$$

Informally, we call a region of the embedding space *dense* when many unrelated identities fall within the decision threshold of a single point—that is, when identity neighborhoods overlap extensively. Our objective is to find points in such dense regions: points whose τ-neighborhood contains as many embeddings as possible. These are masterface embeddings.

4.2. Finding Masterface Embeddings via Genetic Algorithm

Let $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_M\} \subset \mathbb{R}^d$ denote a set of M face embeddings, with $\mathcal{P} \subset \mathcal{E}$. Our goal is to find a point $\mathbf{x} \in \mathbb{R}^d$ that lies within distance τ of as many points in $\mathcal{P}$ as possible:

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathbb{R}^d} \sum_{\mathbf{p} \in \mathcal{P}} \mathbb{I}[d(\mathbf{x}, \mathbf{p}) \leq \tau] \quad \text{s.t.} \quad \|\mathbf{x}\|_2 = 1. \tag{1}$$

Since this objective is non-differentiable and combinatorial, we optimize a smooth surrogate using the LM-MA-ES genetic algorithm [7], which has demonstrated strong performance in high-dimensional optimization [10]. We initialize a population of candidates near the centroid of $\mathcal{P}$, projected onto the unit hypersphere. Each candidate $\mathbf{x}$ is evaluated using:

$$\min_x \mathcal{L}(x) = \frac{1}{|P|} \left[w \cdot C(x) + (1 - w) \cdot D(x) \right], \tag{2}$$

where

$$C(x) = \sum_{p \in P} \mathbb{I}[d(x, p) > \tau], \tag{3}$$

$$D(x) = \sum_{p \in P} d(x, p). \tag{4}$$

$C(x)$ counts embeddings falling *outside* the threshold; minimizing it maximizes coverage. $D(x)$ captures overall proximity, smoothing the optimization landscape. The weight w (default 0.99) prioritizes threshold satisfaction.

A masterface embedding $\hat{\mathbf{x}}^*$ is only the first phase of the attack. To turn it into a practical threat—a face image that can be submitted to a verification pipeline—we apply a second image-space optimization (Phase 2), described in Section 8. The remainder of this section and Sections 6.1–7 analyze the embedding-space Phase 1 attack; Section 8 returns to the pixel domain.

5. Experimental Setup and Headline Results

We evaluate six face-mapping models on the Labeled Faces in the Wild (LFW) dataset [3] (5,749 identities, 13,233 images). Table 1 lists the models and their reproduced LFW accuracy. We test two face detectors (MTCNN [12] and RetinaFace [2]) and two threshold computation directions (FAR$\sim$0.001 from above and from below). Thresholds are computed using the standard LFW 10-fold cross-validation protocol [3]: for each fold we find the FAR$\sim$0.001 threshold and report the average across all ten folds. Distances are measured using the Euclidean metric. All embedding-space searches use the LM-MA-ES genetic algorithm (Section 4.2) with 1,000 generations, population size 100, and loss weight w=0.99. Embedding-space partitioning uses spherical k-means.

For each configuration, we partition the full embedding set (5,749 identities, 1 instance per identity) into k=9 clusters via spherical k-means, run the masterface search independently on each cluster, and evaluate each candidate against the full dataset. We use k=9 throughout to match the masterface budget reported by Shmelkin et al. [10] on the same dataset, enabling direct coverage comparison. We report both per-cluster (best single masterface) and *accumulative* coverage—the fraction of identities within the decision threshold of at least one of the nine masterface points.

Table 2 and Figure 2 present the complete results across all configurations. Two observations are immediate. First, coverage varies dramatically across configurations (from 0.23% to 58.1%) while pairwise accuracy remains tightly clustered between 95% and 99.9%. Second, the CosFace MTCNN-DS row achieves the highest pairwise accuracy in the table (99.65%) and simultaneously the highest accumulative coverage at the strictest threshold (47.2%). The rest of this work dissects this table along its two free axes: threshold direction (Section 6.1) and alignment strategy (Section 6.2). Section 7 then asks what remains when both axes are set conservatively.

6. Security-Critical Pipeline Choices

The headline results above were generated under specific pipeline choices: a particular face-alignment strategy (MTCNN-DS or RetinaFace) and a particular threshold-selection direction (FAR$\sim$0.001 from below or from above). Each choice independently modulates masterface coverage by up to two orders of magnitude—a swing that pairwise accuracy completely fails to flag. We isolate each choice in turn.

6.1. Threshold Selection Sensitivity

The same FAR$\sim$0.001 target, computed from two different directions, produces masterface coverage that differs by up to two orders of magnitude. Table 3 isolates this effect by holding the detector constant (RetinaFace). For CosFace, the "from above" threshold yields 24.6% coverage; the "from below" threshold yields 0.23%—a swing of more than two orders of magnitude under a single rounding choice.

The "from below" and "from above" thresholds differ substantially—by 5–11% of the threshold value (Table 4). The "from below" threshold is strictly conservative, achieving FAR = 0 (zero false accepts across the 10 LFW folds). The "from above" threshold permits FAR $\approx$ 0.003—three times the target. Neither achieves FAR = 0.001 exactly; the true operating point falls somewhere in between.

Each threshold choice represents a different point on the FAR–FRR trade-off. For the angular-margin models (CosFace, ArcFace, AdaFace), choosing the conservative "from below" threshold costs almost nothing in usability: the FRR increase is at most 0.10 percentage points. Yet the "from above" alternative admits coverage that is higher by up to two orders of magnitude. A practitioner selecting the permissive threshold gains negligible usability while unknowingly opening the attack surface by orders of magnitude.

FaceNet presents the opposite pattern. The conservative threshold rejects 12.0% of genuine pairs—nearly double the 6.5% of the permissive threshold. A practitioner choosing "from above" has a strong usability motivation: halving false rejections. Yet even the conservative threshold (FAR = 0, FRR = 12%) admits 41.4% masterface coverage. For FaceNet, the vulnerability is not threshold-sensitive—it is pervasive regardless of the choice.

The core observation is that a single evaluation decision—whether to round the FAR target up or down—determines whether the system admits $<$1% or $>$20% masterface coverage. A practitioner who reports FAR$\sim$0.001 without specifying the computation direction may be operating at a threshold that is up to two orders of magnitude more vulnerable than they realize.

6.2. Alignment Sensitivity

Consider CosFace under MTCNN-DS[1] alignment: 99.65% LFW accuracy, 47.2% masterface coverage. Switch to RetinaFace alignment, holding everything else constant (FAR$\sim$0.001 from below, same model, same matchset, same search): 99.58% accuracy, 0.23% coverage. Accuracy moves 0.07 percentage points; coverage moves by more than two orders of magnitude (from 0.23% to 47.2%). This is the cleanest demonstration in our experiments that pairwise accuracy and masterface coverage are independent properties of a face-verification pipeline. A skeptic might attribute high coverage to degraded accuracy (as in ArcFace under MTCNN-DS, which loses 4 percentage points of accuracy); the CosFace case refutes this directly—accuracy is preserved,

[1]Throughout this section, "MTCNN" refers specifically to the MTCNN-DavidSandberg alignment variant (44px margin, bilinear interpolation, center-weighted face selection), as distinct from the canonical MTCNN.

Model	Acc. (FAR ≤ 0.001)	Source	Architecture
ArcFace	99.52%	[1]	IResNet-100
CosFace	99.58%	[1]	IResNet-100
AdaFace-IR101	99.58%	[4]	IResNet-101
AdaFace-ViT	99.70%	[5]	ViT-Base
FaceNet (CASIA)	97.97%	[9]	InceptionResNet
FaceNet (VGG2)	99.40%	[9]	InceptionResNet

Table 1. Face-mapping models used. Accuracy reproduced via 10-fold LFW evaluation with RetinaFace detection.

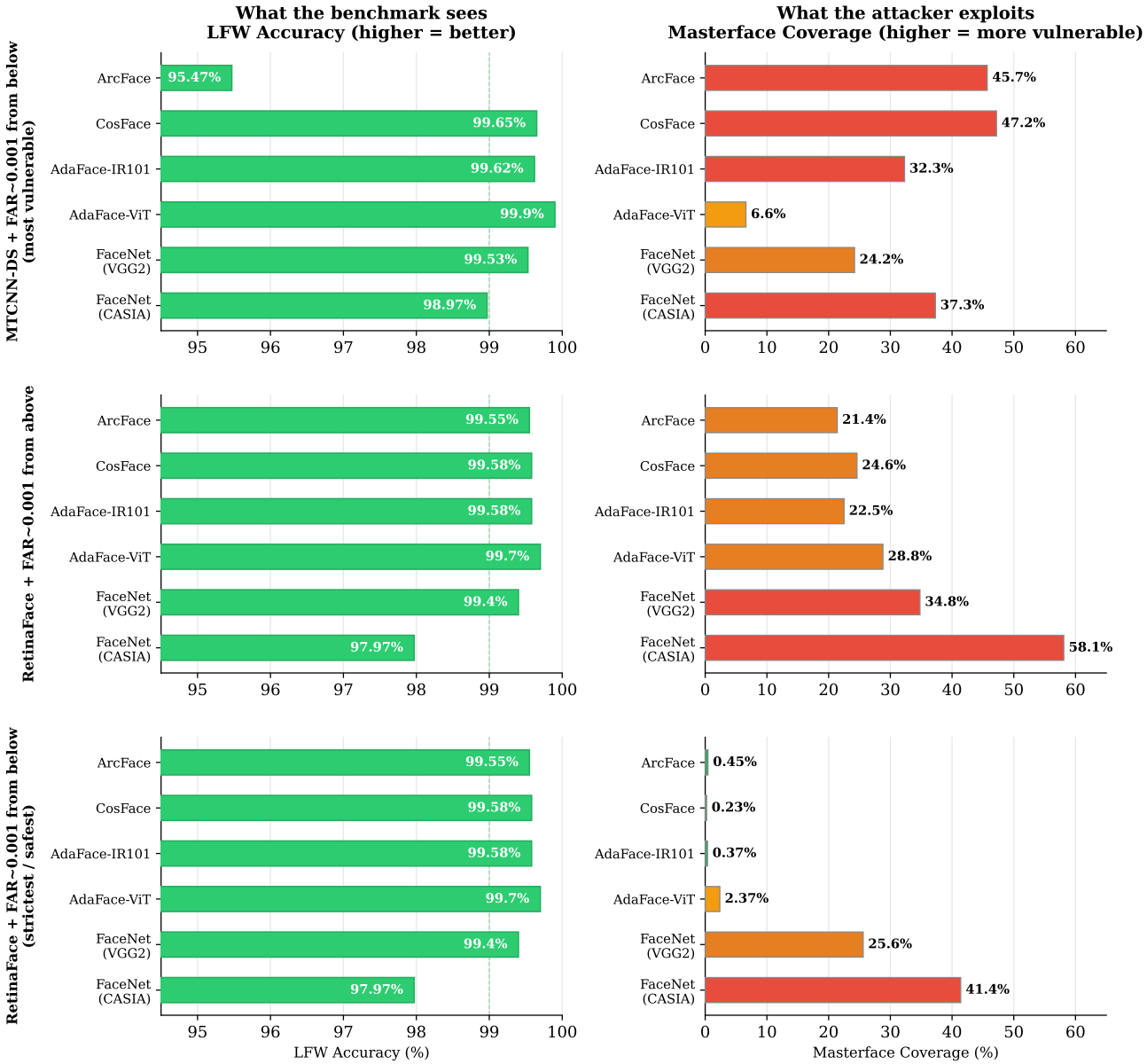

Figure 2. The disconnect between pairwise accuracy and masterface vulnerability across three pipeline configurations, ordered worst (top) to safest (bottom). Left panels: LFW accuracy (higher = better)—uniformly high across all configurations. Right panels: masterface coverage (higher = more vulnerable)—varies by up to two orders of magnitude. Same accuracy numbers, dramatically different attack surfaces. All coverage values are Phase 1 embedding-space coverage across nine masterface embeddings.

Model	Detector / Threshold	Coverage %	Acc. %
ArcFace	MTCNN / under	45.7	95.47
	Retina / over	21.4	99.55
	Retina / under	0.45	99.55
CosFace	MTCNN / under	47.2	99.65
	Retina / over	24.6	99.58
	Retina / under	0.23	99.58
AdaFace-IR101	MTCNN / under	32.3	99.62
	Retina / over	22.5	99.58
	Retina / under	0.37	99.58
AdaFace-ViT	MTCNN / under	6.6	99.90
	Retina / over	28.8	99.70
	Retina / under	2.4	99.70
FaceNet (CASIA)	MTCNN / under	37.3	98.97
	Retina / over	58.1	97.97
	Retina / under	41.4	97.97
FaceNet (VGG2)	MTCNN / under	24.2	99.53
	Retina / over	34.8	99.40
	Retina / under	25.6	99.40

Table 2. Accumulative embedding-space (Phase 1) masterface coverage across nine embeddings (one per cluster). Coverage is the fraction of identities within the decision threshold of at least one masterface point. "Retina" = RetinaFace. "over" = FAR∼0.001 from above; "under" = from below.

Model	Over (%)	Under (%)
CosFace	24.6	0.23
AdaFace-IR101	22.5	0.37
ArcFace	21.4	0.45
AdaFace-ViT	28.8	2.4
FaceNet (CASIA)	58.1	41.4
FaceNet (VGG2)	34.8	25.6

Table 3. Embedding-space coverage with RetinaFace detection: FAR∼0.001 computed from above ("over") vs. from below ("under"). For all angular-margin models, the swing exceeds two orders of magnitude under a single rounding choice; for FaceNet, the swing is comparatively mild.

coverage explodes.

Table 5 extends this comparison to all six models, holding the threshold direction constant at FAR∼0.001 from below.

For the angular-margin IResNet-100/101 models (CosFace, ArcFace, AdaFace-IR101), MTCNN-DS yields coverage higher by up to two orders of magnitude than RetinaFace (e.g., CosFace: 0.23% → 47.2%). For AdaFace-ViT, the same switch is much smaller (2.4% → 6.6%). For both FaceNet variants, the direction reverses: RetinaFace coverage matches or slightly exceeds MTCNN-DS. The two-orders-of-magnitude headline applies specifically to the dominant angular-margin + IResNet combination, not universally to "MTCNN vs. RetinaFace."

What property of the embedding space the alignment switch is reshaping is not established by our experiments. What we can observe directly is that for the affected models the alignment switch leaves pairwise accuracy nearly invariant (CosFace: 0.07 pp; ArcFace: not comparable due to the 4 pp accuracy drop; AdaFace-IR101: 0.04 pp) while coverage moves by up to two orders of magnitude. The alignment effect is a security-critical pipeline decision that pairwise accuracy completely fails to flag.

With the more permissive "from above" threshold, RetinaFace still permits 21.4% (ArcFace) and 24.6% (CosFace) coverage (Table 3). Both the alignment strategy *and* the threshold direction must be chosen conservatively to mitigate the large-scale attack. Whether that is sufficient is the question Section 7 addresses.

7. The Irreducible Floor (JT-Attack)

Sections 6.1 and 6.2 showed that two pipeline choices each modulate masterface coverage by up to two orders of magnitude. They leave open a critical question: does the vulnerability disappear when both choices are set to their safest values? This section addresses that question by shifting from pipeline sensitivity to the models themselves: what remains after the most conservative evaluated pipeline choices are fixed.

The *Joint-Threshold Attack* (JT-Attack) asks: given N *randomly chosen* identities, can we find a single embedding within τ of all of them? All JT-Attack experiments use the strictest pipeline configuration identified above (RetinaFace alignment + FAR∼0.001 from below)—the configuration that drops large-scale masterface coverage below 0.5% for every angular-margin model tested. We sweep N from 2 to 40, running 5 independent repetitions per N value. For each trial, we randomly select N distinct identities from LFW (1 instance per identity) and search for a covering point. To handle cases where the selected embeddings are spatially scattered, we apply spherical k-means with $k = 1, 2, \ldots, 5$ and run the search independently on each partition, evaluating each candidate against the *full* set of N embeddings and keeping the best result across all k values. In a well-separated embedding space—one where identity neighborhoods do not overlap—no such embedding should exist for $N > 1$.

Figure 3 shows the results for ArcFace. The theoretical maximum coverage for a perfect face mapper is $1/N$ (red dashed line)—e.g., 50% for N=2, since one identity's embeddings constitute at most half the set. We observe coverage far exceeding this baseline. For $N = 2, 3, 4$, candidate embeddings satisfying all selected identities are found nearly 100% of the time, even under the strictest pipeline configuration. Even for N=40, coverage remains well above the theoretical maximum. The same pattern holds across all models tested (see Appendix).

Model	$\tau_{\downarrow}$	$\tau_{\uparrow}$	Gap (%)	$\text{FRR}_{\downarrow}$ (%)	$\text{FRR}_{\uparrow}$ (%)
CosFace	1.132	1.262	11.5	0.83	0.80
ArcFace	1.132	1.242	9.7	0.97	0.87
AdaFace-IR101	1.175	1.274	8.4	0.83	0.80
AdaFace-ViT	1.170	1.251	6.9	0.60	0.60
FaceNet (CASIA)	0.958	1.008	5.2	12.0	6.5

Table 4. From-below ($\tau_{\downarrow}$, FAR = 0) and from-above ($\tau_{\uparrow}$, FAR $\approx$ 0.003) thresholds with their false rejection rates. All values averaged over 10 LFW folds using RetinaFace.

	Coverage (%)		Accuracy (%)	
Model	MTCNN-DS	RetinaFace	MTCNN-DS	RetinaFace
CosFace	**47.2**	**0.23**	**99.65**	**99.58**
ArcFace	45.7	0.45	95.47	99.55
AdaFace-IR101	32.3	0.37	99.62	99.58
AdaFace-ViT	6.6	2.4	99.90	99.70
FaceNet (CASIA)	37.3	41.4	98.97	97.97
FaceNet (VGG2)	24.2	25.6	99.53	99.40

Table 5. Embedding-space coverage (nine embeddings) and LFW accuracy with FAR$\sim$0.001 from below: MTCNN-DS vs. RetinaFace, all six models. CosFace (highlighted) is the load-bearing case: accuracy moves by 0.07 percentage points while coverage moves by two orders of magnitude. ArcFace is the one model where alignment also affects accuracy substantially (4 pp drop under MTCNN-DS); for the remaining four angular-margin models accuracy is preserved within 0.3 pp. For both FaceNet variants the coverage effect reverses (RetinaFace coverage $\geq$ MTCNN-DS); for AdaFace-ViT it is weak.

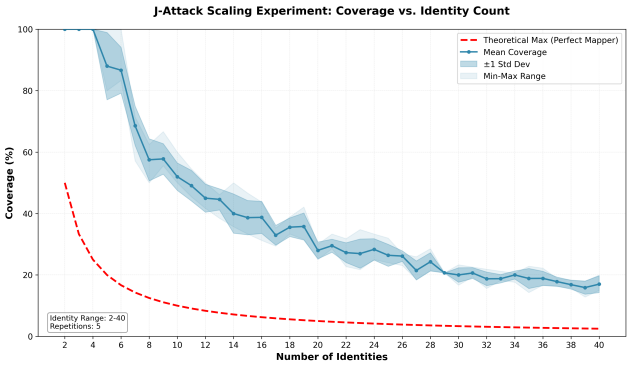

Figure 3. JT-Attack for ArcFace. Coverage (blue) far exceeds the theoretical maximum for a well-separated embedding space (red dashed). For N=2, 3, 4 identities, a single embedding matching all identities is found nearly 100% of the time.

The JT-Attack establishes that pipeline mitigations are bounded from below. Every model tested achieves >98% pairwise accuracy, yet for any 2, 3, or 4 randomly chosen identities, a single point matching all of them is found nearly 100% of the time—even under the safest pipeline we tested. The two pipeline choices in Sections 6.1–6.2 modulate the *magnitude* of large-scale coverage, but they do not eliminate the underlying dense-region structure. That structure lies in the models, not in the pipelines: present across softmax angular-margin losses (ArcFace, CosFace, AdaFace) and triplet losses (FaceNet) alike. Reducing large-scale coverage is a problem the field can solve by reporting and constraining pipeline choices; eliminating the irreducible floor requires changes to the models themselves.

We have shown that the irreducible floor is a property of the embedding space. The remaining question is whether it manifests in the pixel domain—whether these embeddings can be realized as actual face images. Section 8 answers this.

8. From Embeddings to Faces (Phase 2)

The preceding sections established that masterface *embeddings* exist (Section 5), that pipeline choices modulate their large-scale coverage by orders of magnitude (Section 6), and that small-N targeted attacks remain feasible under any pipeline configuration (Section 7). These are all geometric facts about the embedding space. This section returns to the pixel domain and demonstrates that these embeddings translate into plausible face images, making the attack practically realizable.

8.1. Method

Given a masterface embedding $\hat{\mathbf{x}}^*$, we generate a corresponding face image starting from a source face $\mathbf{s}$. We iteratively modify the source so that its embedding approaches $\hat{\mathbf{x}}^*$, inspired by PGD [8]. At each iteration, the loss

$$\mathcal{L}(\mathbf{x}) = d(FM(\mathbf{x}),\ \hat{\mathbf{x}}^*)$$

is minimized using the Adam optimizer [6] applied directly to image pixels, with a perturbation constraint ϵ to preserve

visual similarity to the source face. The result is an image that looks like the source identity but whose embedding matches many unrelated identities. An overview is shown in Figure 4.

8.2. Retention Results

We use a perturbation budget of ϵ=0.25. The attack scenario is concrete: in online identity verification, a user submits a photo (e.g., a digital ID or selfie) that is processed through a face verification pipeline. An attacker who replaces the face region of a legitimate-looking document with a masterface image can pass verification against a large fraction of enrolled identities. The generated image appears as a plausible human face—not a noise pattern or adversarial artifact—making it suitable for systems that perform both visual and algorithmic checks.

Model	Ph1 (%)	Ph2 (%)	Retention
CosFace	47.2	45.7	96.8%
ArcFace	45.7	44.2	96.7%
FaceNet (CASIA)	37.3	34.7	93.0%
AdaFace-IR101	32.3	29.8	92.3%
FaceNet (VGG2)	24.2	21.3	88.0%
AdaFace-ViT	6.6	5.8	87.9%

Table 6. Phase 1 (embedding-space) vs. Phase 2 (image-space) coverage for MTCNN with FAR∼0.001 from below. Retention = Ph2/Ph1.[3] Image generation uses ϵ=0.25 perturbation.

Table 6 shows that under the MTCNN-DS pipeline, 88–97% of embedding-space coverage survives the image generation step. CosFace retains 45.7% of its 47.2% Phase 1 coverage—the generated images are nearly as effective as the raw embeddings. Figure 5 shows the nine generated masterface images for ArcFace; despite appearing as plausible faces, they collectively cover 44.2% of all identities.

The dangerous configuration compounds. CosFace under MTCNN-DS yields both the highest Phase-1 coverage (47.2%) and the highest Phase-1→Phase-2 retention (96.8%). Under RetinaFace pipelines, by contrast, retention drops to 32–55%—e.g., AdaFace-IR101 retains only 7.2%/22.5% = 32% of its Phase-1 coverage when paired with the "from above" threshold. The MTCNN-DS pipeline thus amplifies the threat along both axes simultaneously: it produces more Phase-1 coverage *and* preserves more of it through image synthesis. A practitioner who has not chosen RetinaFace and the conservative threshold faces a doubly amplified attack surface.

The gap between Phase 1 and Phase 2 in any single configuration reflects the difficulty of mapping an arbitrary embedding back to a natural face image within the perturbation budget. Closing this gap—through better image generation methods or larger perturbation budgets—is a direction for future work. However, the core vulnerability is the existence of the embeddings themselves: Phase 1 coverage is the fundamental security metric, and Phase 2 demonstrates that the threat is practically realizable.

[3]For example, CosFace retention: $45.7/47.2 = 96.8\%$; AdaFace-ViT: $5.8/6.6 = 87.9\%$.

8.3. Comparison to Prior Work

Shmelkin et al. [10] reported 43.82% image-space coverage on LFW with FaceNet trained on CASIA-WebFace, using a GAN-based generation method. On the same model and a comparable Phase 2 (image-space) metric, our optimization-based pipeline reaches 49.957% coverage under MTCNN-DS + FAR∼0.001 from above—surpassing their result with a simpler, non-generative method. We emphasize that raw-coverage improvement is not the central contribution of this work; the structural findings of Sections 6–7 are. The Phase 2 comparison is reported here for completeness because it places our method on the same evaluation footing as prior work.

9. Discussion

Open empirical question. Our experiments expose a clear asymmetry in alignment-robustness across models. Angular-margin IResNet-100/101 models (CosFace, ArcFace, AdaFace-IR101) show coverage swings of up to two orders of magnitude across alignment changes. FaceNet (InceptionResNet, triplet loss) and AdaFace-ViT (ViT-Base, angular margin) show much smaller swings—and within AdaFace specifically, the ViT variant is substantially less vulnerable than the IR101 variant under MTCNN-DS (6.6% vs. 32.3%). The embedding spaces produced by the robust models appear more invariant to input-image variation introduced by alignment changes. Whether the distinguishing factor is the architecture (ViT/InceptionResNet vs. IResNet), the training objective (triplet vs. angular margin), or both is left for future work.

Pathways to reduction—and their floor. Our results point at two pipeline-level mitigations for large-scale masterface coverage. First, RetinaFace alignment combined with FAR∼0.001 from below reduces accumulative coverage below 0.5% for every angular-margin IResNet model in our experiments. Second, architecture choice matters: AdaFace-ViT exhibits the lowest large-scale coverage of any model tested under MTCNN-DS (6.6%), and FaceNet variants are largely insensitive to the alignment switch.

All three mitigations reduce large-scale coverage but do not eliminate the JT-Attack floor: under the same strictest pipeline, any 2–4 randomly chosen identities still share a common cover point ∼100% of the time across every model tested. Pipeline mitigations close the door on broad attacks while leaving targeted small-group attacks fully open. The community can solve the first problem by reporting pipeline

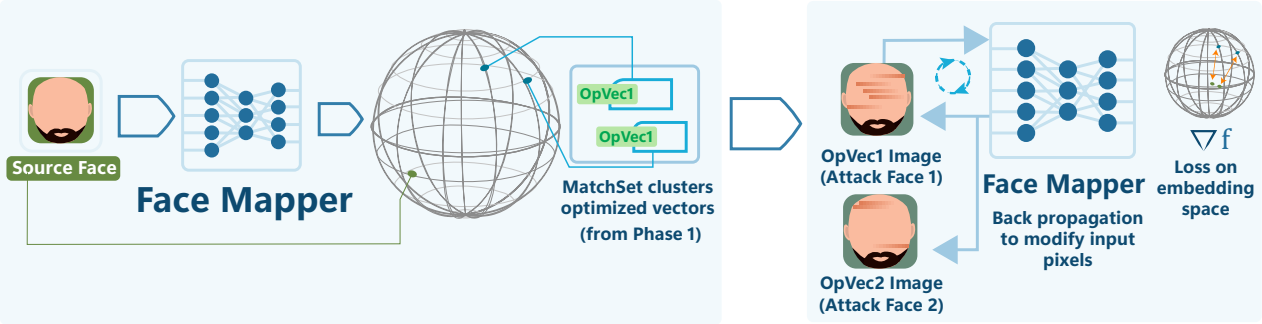

Figure 4. Overview of masterface image generation. A source face is iteratively modified so that its embedding (via the face mapper) approaches the masterface embedding found by the genetic algorithm.

Figure 5. Nine generated masterface images for ArcFace (MTCNN, FAR$\sim$0.001 from below). Top-left: unaltered source face. Each image individually matches hundreds of identities; together they cover 44.2% (Phase 2) of all LFW identities.

choices and adopting coverage metrics; the second requires changes to the models themselves.

Toward adversarial security metrics. We propose that face verification systems report masterface coverage alongside standard accuracy. A practical protocol: run the masterface embedding search (Section 4.2) against the enrolled identity set and report the fraction of identities matched by the best candidate. This provides a direct measure of worst-case dense-region density. The JT-Attack offers a complementary diagnostic: report the maximum N at which a single embedding can match N randomly selected identities with $>$50% success rate. Together, these metrics capture the adversarial security dimension that pairwise accuracy misses.

Limitations. Our experiments are on LFW only; whether the magnitudes reported here transfer to deployed enrollment sets at different scales is untested. Phase 2 results in the main text use a single perturbation budget (ϵ=0.25); other budgets are reported in extended experiments. JT-Attack curves use 5 repetitions per N; statistical confidence intervals would benefit from larger sweeps. Targeted small-group attacks (Appendix A) succeed at 100% for the source/target pairings reported, but extended experiments show that Phase-2 success is sensitive to the source-face identity and perturbation budget—the 100% number characterizes well-chosen pairings rather than worst-case adversaries.

10. Conclusion

We have shown that pairwise verification accuracy is fundamentally insufficient as a security metric. A face verification system can achieve 99.65% pairwise accuracy while 47.2% of its enrolled identities fall within the decision threshold of nine masterface embeddings, and these embeddings can be realized as face images retaining 88–97% of their embedding-space coverage. This is not a failure of one model or one configuration.

This work makes two structural claims. First, two pipeline choices—threshold-rounding direction and face alignment strategy—each modulate masterface coverage by up to two orders of magnitude (e.g., CosFace coverage swings from

0.23% to 47.2% across alignment choices alone). Pairwise accuracy is structurally blind to both. This is an *evaluation failure*: standard benchmarks cannot distinguish a safe configuration from a dangerous one. Second, even when both choices are fixed to the safest setting we tested, the JT-Attack shows that any 2–4 randomly chosen identities still share a common cover point $\sim$100% of the time across every model. This is a *model failure*: current face embedding spaces contain an irreducible floor of dense-region vulnerability that no pipeline choice can configure away.

We propose two changes. (1) Report masterface coverage and JT-Attack curves alongside standard pairwise accuracy—a practical protocol is given in Section 9. This addresses the evaluation failure. (2) Acknowledge that current face embedding models contain an irreducible attack surface for small targeted groups, independent of pipeline configuration. Eliminating this surface requires changes to the models themselves—through training objectives, architectures, or loss functions that constrain dense-region density—not only to how those models are evaluated. The first change is immediate; the second is the open research direction this work points at.

References

[1] DeepInsight. insightface: State-of-the-art 2d&3d face analysis toolbox. https://github.com/deepinsight/insightface. Accessed: 2025-11-23. 5

[2] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 2, 4

[3] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 4

[4] M. Kim. Adaface: Quality adaptive margin for face recognition (github repository). https://github.com/mk-minchul/AdaFace. Accessed: 2025-11-23. 5

[5] M. Kim. Cvlface: A toolkit for face recognition (github repository). https://github.com/mk-minchul/CVLface. Accessed: 2025-11-23. 5

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. 7

[7] I. Loshchilov, T. Glasmachers, and H.-G. Beyer. Large scale black-box optimization by limited-memory matrix adaptation. *IEEE Transactions on Evolutionary Computation*, 23(2):353–358, 2019. 3

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks, 2019. 7

[9] D. Sandberg. Facenet: Face recognition using tensorflow (github repository). https://github.com/davidsandberg/facenet. Accessed: 2025-11-23. 5

[10] R. Shmelkin, L. Wolf, and T. Friedlander. Generating master faces for dictionary attacks with a network-assisted latent space evolution. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 1, 2, 3, 4, 8

[11] M. Wang and W. Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 1, 3

[12] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. In *IEEE Signal Processing Letters*, volume 23, pages 1499–1503. IEEE, 2016. 2, 4

Appendix A: Targeted Small-Group Attacks

Section 7 established that under the strictest pipeline configuration any 2–4 randomly chosen identities still share a common cover point ∼100% of the time. This appendix demonstrates the same effect concretely: we show that a single face image, generated from a chosen source face, can be made to match all four members of a small target group under the strictest configuration.

Source Face	Target Group (4 faces)	Ph1 %	Ph2 %	ϵ
G. Hinton	Beatles	100	100	0.075
Y. LeCun	Beatles	100	100	0.075
Fei-Fei Li	Beatles	100	100	0.075

Table 7. Targeted small-group attack results using ArcFace with RetinaFace and FAR∼0.001 from below—the strictest configuration. A face image of Geoffrey Hinton, with minimal perturbation (ϵ=0.075), matches all four Beatles at 100% success.

Table 7 shows that a face image of Geoffrey Hinton, with perturbation ϵ=0.075, can be made to match all four members of the Beatles with 100% success in both embedding space (Phase 1) and image space (Phase 2). The attack succeeds equally with LeCun and Fei-Fei Li as source faces. We caution that targeted-attack success is sensitive to source-face / target-group pairing: extended experiments with other source faces and a different target group ("scientists") yield Phase-2 success rates ranging from 25% to 100% depending on the pairing and perturbation budget. The Beatles result reported here is representative of cleanly-succeeding pairings; it is not a guarantee that any random source/target combination will reach 100%. The point of this demonstration is that under the strictest pipeline configuration, well-chosen pairings remain trivially attackable—consistent with the JT-Attack finding that the model-level floor admits small-N joint cover points ∼100% of the time. Figure 6 visualizes one successful source-to-target attack.

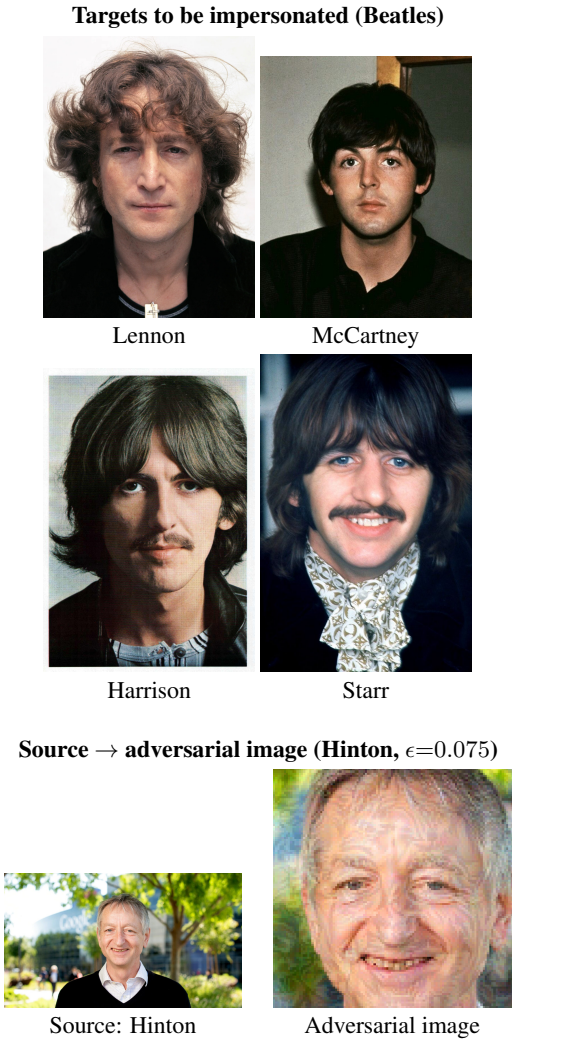

Figure 6. Targeted small-group attack under the strictest configuration (ArcFace + RetinaFace + FAR∼0.001 from below). The top row shows the four Beatles whose embeddings are the impersonation targets. The bottom row shows the source face (Geoffrey Hinton) and the adversarial image generated from it (ϵ=0.075). To a human eye, the adversarial image still looks like Hinton; to the face mapper, its embedding falls within the decision threshold of all four Beatles simultaneously.

Appendix B: JT-Attack Results for Additional Models

Figures 7, 8, 9, 10, and 11 show JT-Attack results for CosFace, FaceNet (both CASIA and VGGFace2 variants), AdaFace-IR101, and AdaFace-ViT. All models exhibit the same pattern: coverage far exceeds the theoretical maximum, confirming that neighborhood overlap is universal across models.

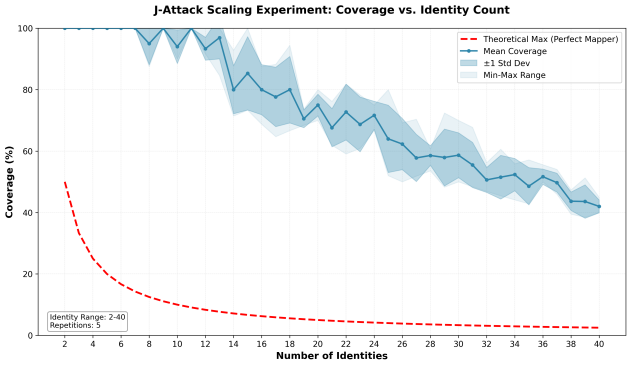

Figure 7. JT-Attack coverage for CosFace.

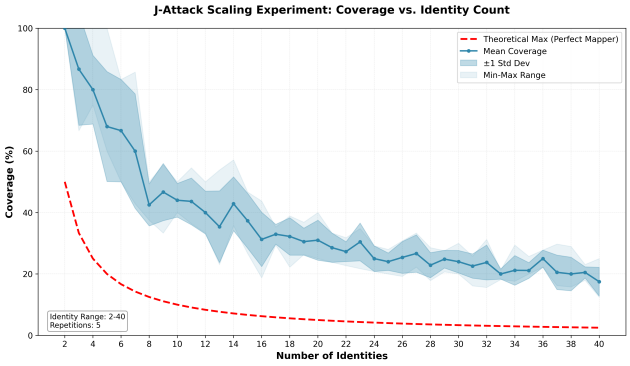

Figure 8. JT-Attack coverage for FaceNet (trained on CASIA).

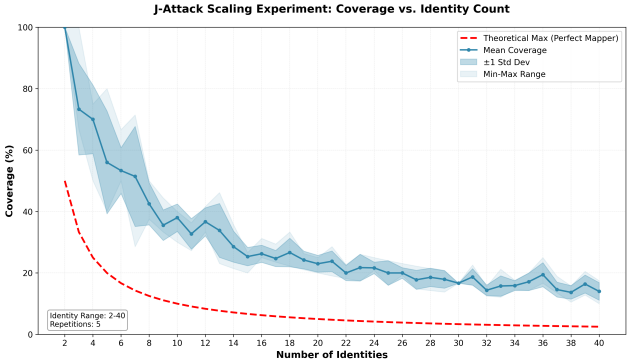

Figure 9. JT-Attack coverage for FaceNet (trained on VGGFace2).

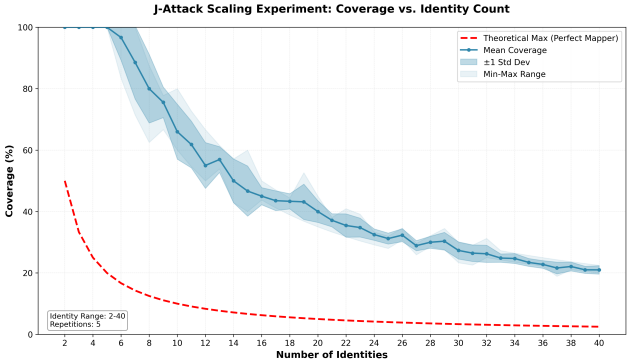

Figure 10. JT-Attack coverage for AdaFace-IR101.

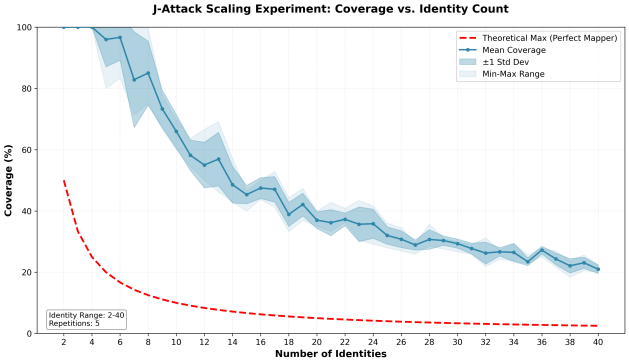

Figure 11. JT-Attack coverage for AdaFace-ViT.