Sketch2ArcFace: Distilling a Robust Sketch Encoder into a Fixed ArcFace Embedding Space from Single Photo–Sketch Pairs

Diego Bonilla Salvador
Independent Researcher
diegobonila@gmail.com

Abstract

We study cross-domain face recognition between unconstrained sketches and photographs when only *single* photo–sketch pairs are available per identity. Our goal is to enable retrieval of sketches against any database of *existing* ArcFace embeddings without re-indexing or re-training that database. To this end, we freeze a pretrained ArcFace model as the target representation and train a sketch encoder to map sketches into the same 512-D hyperspherical space. The training signal blends four ingredients: (i) sampled ArcFace (angular-margin) classification against frozen photo prototypes, (ii) mean-teacher consistency between weak/strong sketch views, (iii) a light CORAL moment-matching penalty that aligns distributional statistics, and (iv) a direct prototype-regression term. Faces in photos are aligned; sketches remain unaligned. We adopt progressive unfreezing of the sketch backbone and an EMA teacher to stabilize learning. Experiments on AP-DrawingDB, CUHK (CUFS/CUFSF), FS2K, WildSketch, and synthetic sketch pairs generated with *FLUX Kontext* show that the proposed recipe yields a sketch encoder whose outputs are immediately compatible with standard ArcFace galleries and can be plugged into downstream diffusion pipelines (e.g., FaceID/IP-Adapter style conditioning) for reconstruction and identity-preserving augmentation. We discuss limitations, ethical considerations, and implications for investigative workflows.

1 Introduction

Large-scale face recognition systems typically operate in a well-curated photographic domain. In practice, however, an investigator, artist, or user might only have an artist's sketch or a stylized rendering of a person. Bridging this cross-domain gap is hard: sketches may differ in pose, proportions, style, and may omit keypoints, making alignment and standard supervised contrastive training unstable when data per identity is scarce. We target the most constrained setting: **one** natural image and **one** sketch per identity, with the additional requirement that the learned representation must remain *drop-in compatible* with any existing ArcFace gallery to avoid re-indexing costs.

Our key idea is to treat the (frozen) ArcFace embedding of the photo as a *class prototype* and distill a sketch encoder into that space using a combination of angular-margin classification with sampled negatives, EMA-based consistency from two augmented views of the sketch, a shallow CORAL penalty, and a direct prototype-regression term. Only photos are aligned to the ArcFace template; sketches are left unaligned to preserve their geometric idiosyncrasies. The outcome is a sketch encoder that "speaks ArcFace" and can query legacy galleries as-is.

Contributions.

- A single-pair-per-identity training protocol that preserves *exact ArcFace geometry* while learning a robust sketch encoder.
- A stabilized objective that combines sampled ArcFace softmax, mean-teacher consistency, CORAL, and prototype regression, with progressive unfreezing and cosine schedule on scale/margin.
- A practical pipeline that aligns photos only, supports retrieval against off-the-shelf ArcFace indices, and in-

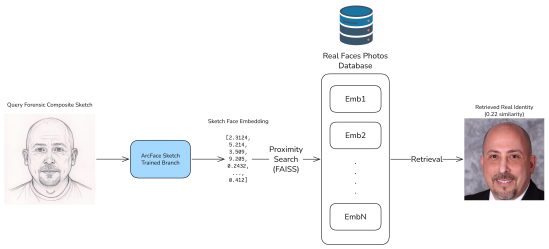

Figure 1: An example retrieval result. On the left is the query sketch; on the right is top-1 retrieved photos from the FFHQ gallery.

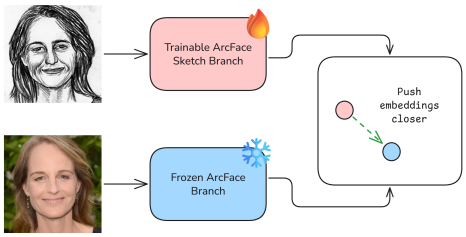

Figure 2: For training, we freeze the real face ArcFace encoder and train a sketch branch to align both embeddings.

tegrates naturally with identity-conditioned diffusion for reconstruction and data augmentation.

2 Related Work

Face embeddings. ArcFace [1] introduced additive angular-margin softmax on the unit hypersphere and is widely deployed through InsightFace [2]. Our work freezes ArcFace and learns a sketch encoder to produce compatible vectors.

Sketch–photo matching. Classical cross-modal face sketch recognition includes CUHK CUFS/CUFSF [10, 11] and subsequent synthesis/matching pipelines (e.g., APDrawingGAN [9]). More recent datasets such as FS2K [12] and WildSketch [13] broaden style diversity and in-the-wild conditions.

Teachers, consistency and domain alignment. We follow the mean-teacher framework [5] for view-consistency and use a shallow CORAL penalty [6] for distribution matching between sketch batches and photo prototypes.

Sampled softmax and mining. To handle many identities with few GPU resources we adopt sampled softmax [7] and mix random with semi-hard negatives as in FaceNet [8].

Backbones. We initialize the sketch encoder from OpenCLIP/CLIP [3, 4], progressively unfreezing deeper blocks during training.

Identity-conditioned diffusion. For reconstruction and augmentation we connect the learned embedding to FaceID-style conditioning (e.g., IP-Adapter FaceID; PhotoMaker personalization) and generate additional sketches with FLUX Kontext models [14, 15, 16, 17].

3 Problem Setting

We are given identities $y \in \{1, \ldots, C\}$ with only one photo p_y and one sketch s_y. Let $f_{\text{arc}} : \mathbb{R}^{H\times W\times 3} \to \mathbb{S}^{d-1}$ be a frozen ArcFace network ($d = 512$), and f_θ a trainable sketch encoder. Define prototypes $v_y = f_{\text{arc}}(A(p_y))$, where A is a face aligner for photos (five-point similarity to 112×112). Sketches are unaligned. The goal is to train f_θ so that $z = f_\theta(x)$ for a sketch x is close (in cosine) to v_y for the correct identity and far from $v_{y'}$, enabling direct search in any ArcFace gallery.

4 Method

4.1 Embedding space and sampled ArcFace softmax

We L2-normalize all embeddings: $\|z\|_2{=}\|v_y\|_2{=}1$. For a mini-batch $\{(x_i, y_i)\}_{i=1}^B$ we draw a subset $\mathcal{C}$ of classes that includes all positives $\{y_i\}$ plus K negatives (mix of random and semi-hard). The logits follow the ArcFace

rule:

$$\ell_{i,c} = \begin{cases} s \cdot \cos(\theta_{i,c}), & c \neq y_i, \\ s \cdot \cos(\theta_{i,y_i} + m), & c = y_i, \end{cases} \tag{1}$$

where $\cos\theta_{i,c} = z_i^\top v_c$, $s > 0$ is a learnable (or scheduled) scale and $m > 0$ is the angular margin. The loss is standard cross-entropy over the sampled set:

$$\mathcal{L}_{\text{arc}} = \frac{1}{B}\sum_{i=1}^{B}\Big[-\log\frac{\exp(\ell_{i,y_i})}{\sum_{c\in\mathcal{C}}\exp(\ell_{i,c})}\Big]. \tag{2}$$

We schedule (s, m) from small values to their targets to avoid early training instability.

4.2 Prototype regression

We directly pull each sketch embedding toward its photo prototype:

$$\mathcal{L}_{\text{reg}} = \frac{1}{B}\sum_{i=1}^{B}\big(1 - z_i^\top v_{y_i}\big). \tag{3}$$

4.3 Mean-teacher consistency

We maintain an EMA teacher θ' updated by $\theta' \leftarrow \tau\theta' + (1-\tau)\theta$. Given two augmentations of the same sketch, weak $x^{(w)}$ and strong $x^{(s)}$, we penalize their mismatch through the teacher prediction:

$$\mathcal{L}_{\text{con}} = \frac{1}{B}\sum_{i=1}^{B}\big(1 - f_\theta(x_i^{(s)})^\top f_{\theta'}(x_i^{(w)})\big). \tag{4}$$

4.4 Shallow CORAL alignment

To keep batch statistics of sketches close to those of photo prototypes, we use the CORAL discrepancy between means and covariances:

$$\mathcal{L}_{\text{coral}} = \|\mu_z - \mu_v\|_2^2 + \|\Sigma_z - \Sigma_v\|_F^2, \tag{5}$$

where (μ_z, Σ_z) are the batch mean/covariance of sketch embeddings and (μ_v, Σ_v) are computed on a random prototype sample of comparable size.

4.5 Total objective and schedules

The full loss is

$$\mathcal{L} = \mathcal{L}_{\text{arc}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{con}}(t)\mathcal{L}_{\text{con}} + \lambda_{\text{coral}}\mathcal{L}_{\text{coral}}, \tag{6}$$

with $\lambda_{\text{con}}(t)$ decayed after backbone unfreezing to prevent consistency domination. We also use cosine LR with warmup; the ArcFace scale/margin (s, m) are cosine-ramped from $(16, 0)$ to $(64, 0.3)$.

4.6 Architectural and training details

Backbone. We start from OpenCLIP ViT-B/16 [4, 3] and train a two-layer head to 512-D with L2 normalization. *Progressive unfreezing:* the head is trained first; after a short warmup we unfreeze the last k transformer blocks.

EMA teacher. The teacher is an EMA of the student with decay τ ramped from 0.996 to 0.999 across epochs [5].

Negatives. For the sampled-softmax we include all positives and fill with a 50–50 mix of random and semi-hard negatives (top-k by similarity, excluding the positive) [7, 8].

Alignment and augmentations. Photos are aligned to the ArcFace 112×112 template using InsightFace's five-point landmarks. Sketches are *not* aligned; we apply weak/strong augmentations (resized crops, light affine/perspective jitter, flips, blur/sharpness, slight contrast, occasional inversion) to simulate style and stroke variability [2].

Optimization. We use AdamW with gradient clipping and gradient accumulation for large effective batch sizes. A new backbone param group is added *without* resetting optimizer state when unfreezing to preserve momentum.

5 Datasets

We train on the union of: (i) **APDrawingDB** [9] (photo–artist portrait pairs), (ii) **CUHK** CUFS/CUFSF [10, 11], (iii) **FS2K** [12] (2,000 high-quality face sketches with

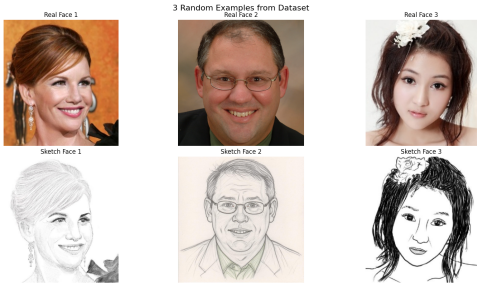

Figure 3: Three random examples from our paired dataset.

annotations), (iv) **WildSketch** [13] (in-the-wild sketches). In addition we synthesize sketch or stylized pairs from photos using the **FLUX Kontext** family [17, 16], creating multiple styles per identity while keeping a single ground-truth identity. Final dataset contains 4,872 pairs.

Single-pair limitation. Each identity contributes one photo and one sketch. We address this by: (1) *prototype supervision*—the frozen ArcFace photo embedding serves as a stable target; (2) *two-view consistency* on a single sketch; (3) *distribution matching* via CORAL over mini-batches; and (4) *semi-hard negatives* drawn from many other identities, which supplies contrastive signal even without multiple positives.

6 Evaluation Protocol

Retrieval. Given a sketch query, we compute its embedding z and rank by cosine similarity against either (i) a *val-only* photo-prototype gallery or (ii) a *full* gallery that includes both train and val identities (more realistic at scale). We report Recall@1 for both.

Verification. For each sketch we compute its positive score $z^\top v_y$. We sample K impostor scores against random gallery identities and estimate TPR at fixed FPR (10^{-2}, 10^{-3}).

7 Results and Analysis

In our internal runs (single-pair-per-ID, OpenCLIP ViT-B/16, ArcFace prototypes frozen), the model learns a stable embedding compatible with ArcFace galleries. We observe that (i) gradually increasing (s, m) avoids early collapse, (ii) unfreezing a few final blocks improves high-recall behavior, and (iii) the mean-teacher term is most useful before unfreezing; decaying it afterward prevents over-regularization.

7.1 Evaluation Metrics and Results

What we measure. We evaluate two complementary tasks: (i) retrieval, where a query sketch must retrieve its corresponding face from a gallery, and (ii) verification, where we decide if a sketch and a photo belong to the same identity. We also report the optimization losses and learning-rate schedules used during training.

Metric definitions.

- **Recall@1 (Full Gallery)**: Fraction of queries whose correct match is ranked first when searching the entire gallery (hardest setting).
- **Recall@1 (Validation Gallery)**: Same as above but restricted to a held-out validation gallery; used for model selection.
- **TPR@FPR=10^{-2} and 10^{-3}**: Verification performance measured as the true positive rate at operating points where false positive rate is 10^{-2} and 10^{-3}, respectively. Higher is better; lower FPR is stricter.
- **Losses** (lower is better): $\mathcal{L}_{\text{CE}}$ is identity classification cross-entropy; $\mathcal{L}_{\text{cons}}$ encourages sketch–photo embedding consistency; $\mathcal{L}_{\text{reg}}$ regresses toward target face embeddings (e.g., L1/Huber); $\mathcal{L}_{\text{coral}}$ aligns second-order statistics across modalities (CORAL). The overall training loss is a weighted sum of these terms.
- **Learning rates**: separate schedules for the backbone ($\text{lr}_{\text{backbone}}$) and the task head ($\text{lr}_{\text{head}}$).

Final evaluation. Training ran for 50 epochs. Tables 1–2 report best and final metrics.

Training schedule. lr_{head}: initial 1.00×10^{-3}, final 5.45×10^{-5}, min 2.00×10^{-5}. $\text{lr}_{\text{backbone}}$: initial 3.00×10^{-5}, final 1.63×10^{-6}, min 6.00×10^{-7}.

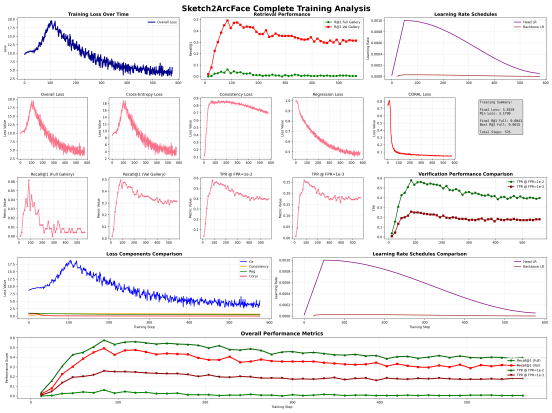

Figure 4: Summary of all training metrics.

Table 1: Retrieval and verification (higher is better).

Metric	Best	Final
Recall@1 (Full Gallery)	0.0615	0.0041
Recall@1 (Validation)	0.4918	0.3156
TPR @ FPR=10^{-2}	0.5779	0.3975
TPR @ FPR=10^{-3}	0.2582	0.1803

Discussion. Validation retrieval peaks at Recall@1 $= 0.4918$, with verification up to TPR $= 0.5779$ at FPR $= 10^{-2}$. Performance on the full gallery is substantially harder (best Recall@1 $= 0.0615$), reflecting the larger search space and domain gap. Final-epoch scores are lower than the best, suggesting that early stopping on validation metrics would yield stronger reported results.

Ablations. Removing the consistency term slows convergence; removing prototype regression harms identity fidelity; removing CORAL slightly degrades cross-style robustness. Hard-negative mixing boosts Recall@1 under large galleries.

8 Applications

Retrieval against legacy ArcFace indices. Because our encoder outputs remain in the ArcFace space, any existing database of ArcFace embeddings (e.g., created with InsightFace pipelines) can be searched without re-indexing or model upgrades [2, 1].

Table 2: Training losses (lower is better).

Loss	Final	Min	Mean
Overall	5.8159	3.1790	9.1765
Cross-Entropy	5.3032	2.6713	—
Consistency	0.6957	0.1101	—
CORAL	0.0455	0.0411	—
Regression	0.4986	0.4553	—

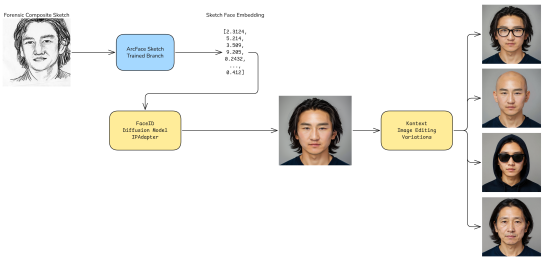

Figure 5: Face reconstruction from FaceID method and automatic variations using FLUX Kontext. All from a single face sketch input.

Reconstruction and augmentation. The sketch embedding can condition a diffusion model via a FaceID/IP-Adapter pathway to reconstruct a photorealistic face, after which FLUX Kontext generates diverse identity-preserving stylizations for analysis or data enrichment [14, 15, 17].

Implications for investigations. The approach supports querying a suspect sketch against large photographic galleries and generating controlled reconstructions to solicit witness feedback. We caution that performance degrades under extreme abstraction, missing facial regions, or adversarial sketches; human oversight and legal safeguards remain essential.

9 Ethical Considerations

This work deals with biometric data. Any deployment must comply with applicable law and policy, ensure appropriate consent, audit for bias across demographic groups, and prohibit uses that violate civil liberties. Synthetic augmentation must not be mistaken for ground truth.

10 Limitations and Future Work

We still rely on high-quality five-point alignment for photos and on the availability of at least one photo per identity to anchor the prototype. Future directions include multi-scale sketch features, cross-attention with photo priors, and lightweight on-the-fly negative caching for larger galleries.

11 Conclusion

We presented a practical recipe to map unconstrained face sketches into a fixed ArcFace space using only single photo–sketch pairs per identity. By combining sampled ArcFace softmax, mean-teacher consistency, CORAL, prototype regression, and progressive unfreezing, we obtain an encoder that is directly compatible with legacy ArcFace indices and plays well with identity-conditioned diffusion for reconstruction and augmentation.

Reproducibility. The training script includes all components described here (alignment, schedules, optimizer state management when unfreezing, EMA teacher, evaluation).

References

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *CVPR*, 2019.

[2] InsightFace Team. InsightFace: 2D and 3D Face Analysis Library. `https://insightface.ai/`, accessed 2025.

[3] A. Radford et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

[4] G. Ilharco et al. OpenCLIP. 2021–2023. `https://github.com/mlfoundations/open_clip`.

[5] A. Tarvainen and H. Valpola. Mean Teachers are Better Role Models: Weight-averaged Consistency Targets Improve Semi-supervised Deep Learning Results. In *NeurIPS*, 2017.

[6] B. Sun and K. Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *ECCV Workshops*, 2016.

[7] S. Jean, K. Cho, R. Memisevic, and Y. Bengio. On Using Very Large Target Vocabulary for Neural Machine Translation. In *ACL*, 2015. (Sampled Softmax)

[8] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *CVPR*, 2015.

[9] S. Xiao et al. APDrawingGAN: Generating Artistic Portrait Drawings from Face Photos with Hierarchical GANs. In *CVPR*, 2019. (APDrawingDB)

[10] S. Zhang, R. Ji, C. Li, B. Zhang, Q. Tian. Coupled Information-Theoretic Encoding for Cross-Modal Matching and Generation. In *CVPR*, 2019. (CUFS).

[11] X. Wang and X. Tang. Face photo–sketch synthesis and recognition. *IEEE TPAMI*, 31(11):1955–1967, 2009. (CUFSF)

[12] X. Chen et al. FS2K: High-Quality Face Sketch Synthesis with Facial Component Prior. *arXiv:2208.08728*, 2022.

[13] C. Chen et al. WildSketch: Face Sketch-to-Cartoon Translation with Unpaired Training. In *ICCV Workshops*, 2021/2022.

[14] Z. Ye et al. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *arXiv:2308.06721*, 2023. (FaceID variants)

[15] Z. Liang et al. PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding. *arXiv:2312.04461*, 2023.

[16] BFL. FLUX 1.1 Models (Model Cards and Documentation). 2024. `https://blackforestlabs.ai/flux-1-1/`.

[17] Flux AI. Flux.1 Kontext Image Generation and Editing. 2024–2025. `https://flux-ai.io/flux-kontext/`.