

Identity-Preserving Sketch-to-Photo Synthesis for Forensic Face Retrieval Using Discriminator Regularization

Nuntipat Phisutthangkoon¹, Sirapat Thianphan^{1*} and Thiansiri Luangwilai²

¹*Department of Computer Science, Navaminda Kasatriyadhiraj Royal Air Force Academy, Thailand*

²*Mathematics Department, Navaminda Kasatriyadhiraj Royal Air Force Academy, Thailand*

Received: 27 May 2026, Received in revised form: 11 September 2026, Accepted: 24 September 2026

Available online: 7 October 2026

Abstract

Background and Objectives: Forensic investigations often begin with a hand-drawn sketch, and the practical question is not whether it can be turned into a plausible face but whether that face retrieves the right person from a gallery of photographs. CycleGAN has become the standard route because it trains without paired supervision. Evaluation of such models, however, leans on reconstruction measures such as SSIM and PSNR, which reward smooth, photograph-like output; whether a face scoring well on them also carries the identity information a recognizer depends on is rarely tested. A second gap concerns the discriminator, where regularization is treated as a matter of optimization stability and its effect on identity-discriminative structure is left unexamined. Our objective was to determine whether regularizing the CycleGAN discriminator with all-block dropout, on top of spectral normalization, improves identity-preserving retrieval, whether generator depth or dropout rate contributes independently, and whether pixel-level fidelity can serve as a proxy for retrieval when a model is selected for forensic use.

Methodology: Experiments used the CUHK student subset of CUFS, 188 sketch-photo identity pairs at 128×128 resolution, partitioned under fixed seeds into five identity-disjoint splits of 131 training, 20 validation, and 37 test identities. Six configurations were trained on every split, giving thirty runs: two spectral-normalization-only baselines with ResNet-6 and ResNet-9 generators, and four configurations adding all-block discriminator dropout at rates of 0.3 and 0.5 for each depth. Both baselines are previously published spectral-normalized architectures, retrained here so that all configurations share identical partitions, checkpoints, and criteria. Training ran 200 epochs with the standard CycleGAN objective, with cycle and identity loss weights held fixed so that discriminator regularization was the only variable. Evaluation ran in two stages that never share data: checkpoints saved every twenty epochs from epoch 100 were scored on the validation gallery, the best validation Top-1 was selected with ties resolved by a rule fixed in advance, and that checkpoint was evaluated once on the held-out test identities. Synthesized photographs were embedded with a fixed InsightFace ArcFace recognizer, ranked by cosine similarity, and scored

as Top-1, Top-5, and Top-10 accuracy alongside SSIM and PSNR. Retrieval is reported under two denominators: a conditional figure over identities whose query and gallery counterpart were both embedded, and an end-to-end figure over all 37 test identities counting a query as incorrect whenever any pipeline stage fails. Paired Wilcoxon signed-rank tests across the five splits are descriptive, since five paired observations floor the attainable p -value at 0.0625 and the test sets overlap.

Main Results: Every configuration carrying all-block dropout outperformed both baselines on conditional Top-1, on all five splits without exception; end-to-end, the same holds in thirty-nine of the forty configuration–split comparisons, with one tie and no reversal. Holding depth fixed, dropout raised mean conditional Top-1 by 14.6 percentage points at ResNet-9, from 80.14% to 94.74%, and by 8.0 points at ResNet-6, from 83.17% to 91.16%, both at the minimum attainable p -value of 0.0625; end-to-end the gaps are 13.0 and 6.5 points. Neither depth nor rate produced a comparable effect: the baselines differed by roughly three points, and no depth or rate comparison held a consistent direction across splits, pointing to the presence of regularization rather than its strength or generator size as the operative factor. The synthesis measures inverted this ordering. The ResNet-9 baseline achieved the highest SSIM, 0.6930, and the highest PSNR, 18.23 dB, while returning the lowest Top-1 of the six. Across all thirty method–split combinations, mean SSIM and Top-1 were negatively correlated (Pearson $r = -0.55$, $p = 0.002$), and at the level of individual images the two were effectively unrelated. Training curves are consistent with discriminator overfitting in the unregularized baselines, which the dropout runs avoid.

Conclusions: Crossing generator depth with discriminator dropout, all-block dropout in the discriminator consistently improves identity-preserving retrieval in sketch-to-photo synthesis, and the benefit comes from the presence of the regularization rather than a particular rate or depth. The more consequential finding is that pixel fidelity did not predict retrieval and can point the opposite way: the configuration reconstructing most faithfully retrieves worst, so a model chosen on SSIM or PSNR would be among the weakest search tools here. For forensic systems, where the deliverable is a correct ranked identity rather than a faithful portrait, synthesis models should be selected and stopped on retrieval measured against a held-out gallery. Results come from a single small, clean dataset with five splits, and confirmation on larger, noisier forensic-like data remains necessary.

Keywords: forensic face retrieval; sketch-to-photo synthesis; CycleGAN; discriminator regularization

Introduction

In real-world forensic investigations, a hand-drawn or composite facial sketch may be the only usable representation of a suspect, particularly when surveillance footage is unavailable or unsuitable for reliable identification (Chugh *et al.*, 2017; Tang & Wang, 2003). Such sketches are usually created from eyewitness descriptions and used to search galleries of mugshots or identity records for possible matches. As shown in Figure 1, the goal is to use a sketch as a query and retrieve the correct person from a ranked list of face photographs. This task is challenging because sketches and photographs differ substantially in visual characteristics. Sketches emphasize contours and salient facial structures while suppressing photometric cues. As a result, the sketch-photo modality gap makes direct matching unreliable and can lead to unstable retrieval performance, especially when sketch quality varies with witness memory or drawing style (Guo *et al.*, 2020).

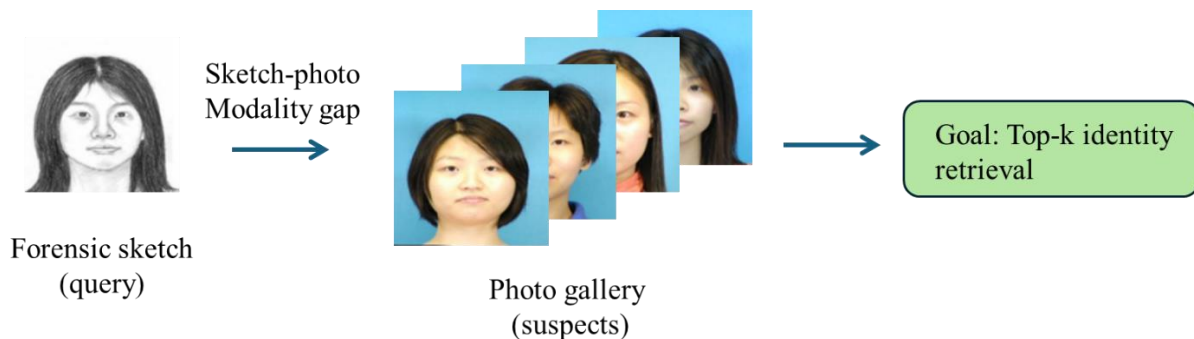


Figure 1 Real-world forensic retrieval scenario. A hand-drawn forensic sketch is used as a query to search a gallery of face photos, aiming to return the correct identity among the Top-*k* ranked candidates.

A practical way to reduce this gap is to translate one modality into the other before retrieval or recognition (Wang *et al.*, 2017; Tang & Wang, 2003). Instead of matching sketches directly to photographs, the system first synthesizes a photo-like face and then retrieves within a single domain, that is, photo-to-photo matching. This suits forensic applications because existing photo-based recognition systems can be reused without redesigning them for cross-modal input. We adopt a retrieval-oriented framework in which CycleGAN performs sketch-to-photo synthesis, followed by identity retrieval using a fixed ArcFace pipeline with cosine-similarity ranking. CycleGAN learns the sketch-photo mapping through cycle consistency, while ArcFace provides a discriminative embedding space for ranking candidate identities. ArcFace is kept fixed so that retrieval performance reflects the usefulness of the synthesized images rather than recognizer adaptation (Deng *et al.*, 2019).

Existing work has largely followed two directions: translating between sketch and photo domains for recognition or retrieval, and improving identity preservation so synthesized images remain reliable for matching (Fang *et al.*, 2020; Guo *et al.*, 2020; Wang *et al.*, 2017). Phisutthangkoon *et al.* (2026) showed that stabilizing CycleGAN with spectral normalization and increasing generator depth can improve CUFS synthesis fidelity in SSIM and PSNR, measures widely used in image restoration and synthesis (Wang *et al.*, 2004). Later studies pursued stronger global modeling and cross-domain representation learning to capture long-range and hierarchical facial cues (Du *et al.*, 2024). Diffusion models have since been applied to the same problem: Que *et al.* (2024) synthesize photographs through sequential denoising and report improved recognition accuracy, whereas Tang *et al.* (2025) correct latent features to preserve identity. Both frame the challenge as achieving visual quality and identity preservation at once, though neither is directly comparable here given differences in recognizer, dataset, and protocol. High-resolution GAN frameworks have likewise introduced perceptual, style, and detail-aware losses to retain fine facial structure under their own recognition settings (Tang *et al.*, 2024; Mahfoud *et al.*, 2024; Huang *et al.*, 2024), while identity-aware translation adds recognition supervision to favor identity consistency over realism alone (Fang *et al.*, 2020), and detail-preserving methods target fine facial structure during synthesis (Wan *et al.*, 2021). Taken together, these studies indicate that visual fidelity matters but does not suffice; forensic retrieval depends more on identity-preserving synthesis judged with Top-*k* accuracy and Cumulative Match Characteristic (CMC) curve. Direct numerical comparison remains difficult, since these works differ substantially in recognizer design, supervision, dataset, and evaluation protocol. Of them, Phisutthangkoon *et al.* (2026) is the most directly relevant controlled baseline for the present work, whereas Du *et al.* (2024) and Tang *et al.* (2024) are better read as contextual references than as like-for-like benchmarks under the fixed ArcFace pipeline adopted here.

A synthesized image may look plausible yet distort the identity cues a face embedding relies on, pushing the correct subject down the ranking or confusing it with similar candidates. We therefore treat sketch-to-photo synthesis as a retrieval problem rather than a pure image-generation task, and alongside SSIM and PSNR we emphasize the retrieval measures standard in sketch-based face recognition: Top-1, Top-5, and Top-10 accuracy with the CMC curve (Mahfoud *et al.*, 2022; Chugh *et al.*, 2017). ArcFace retrieval also requires successful face detection and alignment, which some synthesized images fail through artifacts or weak facial structure; to avoid overstating usability we report embedding success separately for queries and gallery photographs, and give retrieval over all test identities as well as over those that embed.

To improve identity preservation we regularize the CycleGAN discriminator, combining spectral normalization with all-block dropout applied after each intermediate block but not the output layer, aiming to curb discriminator overfitting under limited training identities and to generalize to unseen subjects and sketch variations.

Rather than a single train–test split, we evaluate every configuration across five random identity-disjoint splits of CUFS under a closed-set protocol in which the query is the photo synthesized from a test sketch and the gallery holds the real test photos, so that the reported differences reflect the regularization rather than a favorable partition.

Methodology

1. *Overall Workflow of the Study*

Retrieving a suspect from a photo gallery using only a hand-drawn sketch requires bridging two modalities that share little at the pixel level. The approach taken here is to close that gap by synthesis: a sketch is first translated into a photo-like image, and retrieval is then carried out entirely within the photo domain, where a face recognizer can operate as it normally would. What distinguishes this study is where the intervention is placed. Rather than modifying the generator or the recognizer, we regularize the discriminators that supervise the generator during training, and we measure the consequence not by how faithfully the output reconstructs a photograph but by whether it retrieves the correct identity.

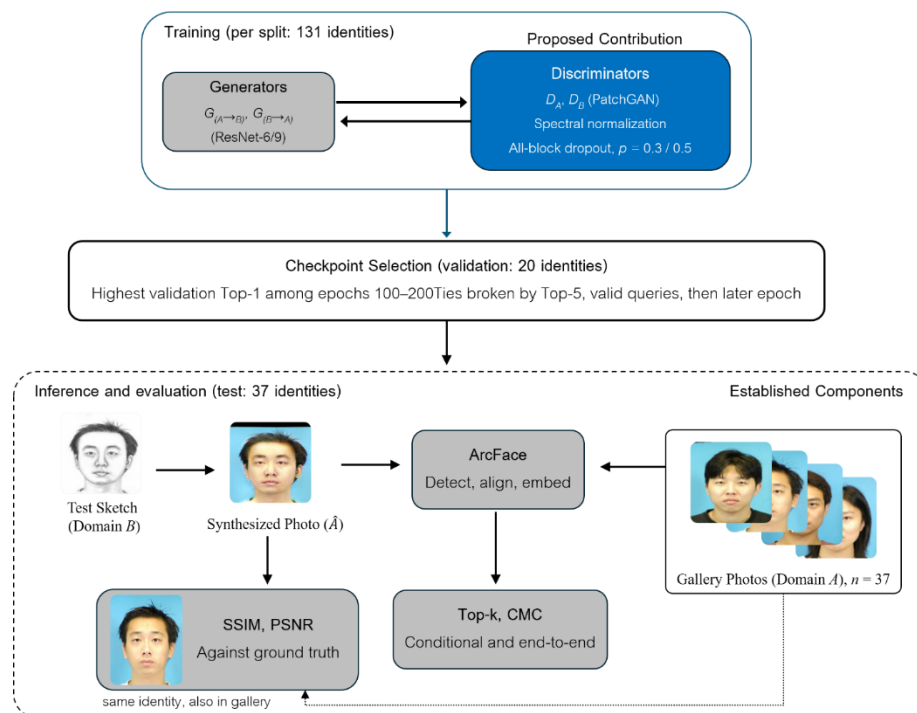


Figure 2 Sketch-to-photo retrieval framework; the discriminator components examined here are shown in blue.

Six configurations, all with spectral normalization and with or without all-block dropout, are evaluated across five identity-disjoint CUFS splits (131/20/37).

Figure 2 summarizes the resulting framework, separating the components introduced in this work from the established ones it builds on. Training and evaluation are carried out independently on each of five identity-disjoint splits of the CUFS dataset, partitioned into 131 training, 20 validation, and 37 test identities. For each split, a CycleGAN model is trained to map Domain B to Domain A , with the regularization examined in this study applied to the two discriminators, D_A and D_B , during training only. Checkpoints saved every twenty epochs from epoch 100 are scored on the validation identities, and the epoch with the highest validation Top-1 is carried forward; the test identities are never consulted at this stage.

At inference, a test sketch from Domain B serves as the query image, and the selected generator $G_{B \rightarrow A}$ performs sketch-to-photo synthesis to produce a photo-like image, denoted as $\hat{A}$. The synthesized query is then processed by a fixed ArcFace pipeline consisting of face detection, alignment, and embedding extraction. In parallel, the 37 real gallery photos from the test set in Domain A are processed by the same pipeline to obtain their embeddings. Retrieval is performed by computing cosine similarity between the query embedding and each gallery embedding, followed by ranking the gallery images in descending order of similarity. The ranked results are used to compute Top-1, Top-5, and Top-10 accuracy, as well as the CMC curve.

In addition to retrieval performance, synthesis quality is evaluated by comparing each synthesized image with its corresponding ground-truth photo using SSIM and PSNR. Retrieval performance is therefore reported together with embedding success for queries and gallery photographs, and over all test identities as well as over those that embed. Applying this protocol consistently across all five splits provides a more reliable assessment of retrieval-oriented sketch-to-photo synthesis than a single partition would allow.

2. *Dataset, Preprocessing, and Experimental Protocol*

This study uses the CUHK Face Sketch Database (CUFS), which contains paired face images of the same identity in two modalities: a hand-drawn sketch and a corresponding photo. This paired structure supports both sketch-to-photo translation and objective evaluation, since each synthesized output can be compared directly with its ground-truth photo. In this paper, the sketch domain is denoted as Domain B and the photo domain as Domain A . CUFS is treated as an identity-based dataset, where each identity contributes one sketch-photo pair, matching the forensic scenario in which a sketch-derived query is used to retrieve the correct identity from a photo gallery. Experiments use the CUHK student subset of CUFS, comprising 188 identities, each with one sketch and one photo; all 188 are used, with no identity filtered out on the basis of image quality or any other criterion.

To reduce sensitivity to a single data partition, we conduct experiments on five random identity-disjoint splits of the dataset. In each split, 131 identities are used for training, 20 for validation, and 37 for testing, with no identity appearing in more than one subset. The five splits were generated with fixed random seeds of 42, 123,

2024, 7, and 99, and the identity membership of every subset is provided in the repository listed under Data Availability. Because the five splits are drawn from the same pool of 188 identities, an identity may appear in more than one test set: the ten pairs of splits share between 2 and 12 test identities, averaging 7.3, which matches the 7.3 expected under independent sampling ($37^2 / 188$). Of the 188 identities, 60 never enter a test set, 85 enter one, 31 enter two, 10 enter three and 2 enter four. The validation subset serves only to select the training checkpoint for each configuration, as described in Section 5, and is never used to report performance. Under the closed-set retrieval protocol, the query is the synthesized photo ($\hat{A}$) generated from a test sketch ($B \rightarrow A$), while the gallery consists of the 37 real test photos in Domain A. Retrieval is performed by ranking gallery identities according to similarity between the query and gallery embeddings, producing a Top- k candidate list.

All images are resized to 128×128 pixels for consistent CycleGAN training and evaluation. Synthesis quality is measured using SSIM and PSNR between each synthesized photo and its corresponding ground-truth test photo. Retrieval performance is evaluated across the five splits using valid-query counts, Top-1, Top-5, Top-10, and CMC, with all results reported as the mean and sample standard deviation over the five splits. Figure 3 shows an example sketch-photo pair from the CUFS dataset.



Figure 3 Example of a sketch (Domain B) and its corresponding photo (Domain A) from the CUFS dataset.

3. *Sketch-to-Photo Synthesis Model and Training*

We perform sketch-to-photo synthesis using a CycleGAN framework that learns a mapping from the sketch domain (Domain B) to the photo domain (Domain A). The model follows the standard CycleGAN architecture with two generators, $G_{A \rightarrow B}$ and $G_{B \rightarrow A}$, and two discriminators, D_A and D_B . During inference, the generator $G_{B \rightarrow A}$ translates a test sketch into a photo-like image, denoted as $\hat{A}$, which is then used as the retrieval query. Both discriminators follow a PatchGAN design to encourage locally realistic synthesis.

3.1 *Architecture and Configurations*

The generator is a ResNet-based CycleGAN generator, and two generator depths are evaluated by varying the number of residual blocks: ResNet-6 and ResNet-9. Both depths are included because generator

capacity is a plausible alternative explanation for any change in retrieval performance; evaluating each regularization setting at both depths allows the effect of discriminator regularization to be separated from that of generator size rather than assumed. The discriminator is a PatchGAN discriminator with spectral normalization applied to its convolutional layers for improved training stability. In the proposed settings, discriminator regularization is further enhanced by all-block dropout, which is applied after the activation in Blocks 1-4, while no dropout is used in the output layer. Six configurations follow from crossing the two generator depths with three discriminator settings: spectral normalization alone, and spectral normalization combined with all-block dropout at $p = 0.3$ and $p = 0.5$. Table 1 lists them in that order, beginning with the two spectral-normalization-only configurations, R6-Base and R9-Base, which use the architectures introduced by Phisutthangkoon *et al.* (2026). Because that study used a different data partition, we retrained and re-evaluated both from scratch under the unified five-split protocol adopted here, so that all configurations are compared on identical partitions, checkpoints, and evaluation criteria. Including both depths without dropout gives every proposed configuration a baseline that differs from it in exactly one factor.

Table 1 CycleGAN model settings for sketch-to-photo synthesis ($B \rightarrow A$). Configurations are ordered by experimental design; the results tables order the same configurations by retrieval performance

Method	Generator (ResNet Blocks)	Discriminator regularization
R6-Base	ResNet-6	Spectral normalization only
R9-Base	ResNet-9	Spectral normalization only
R6-Drop30	ResNet-6	Spectral normalization; all-block dropout ($p = 0.3$)
R9-Drop30	ResNet-9	Spectral normalization; all-block dropout ($p = 0.3$)
R6-Drop50	ResNet-6	Spectral normalization; all-block dropout ($p = 0.5$)
R9-Drop50	ResNet-9	Spectral normalization; all-block dropout ($p = 0.5$)

Note: All-block dropout is applied after discriminator Blocks 1–4, with no dropout in the output layer

3.2 Loss Functions and Training Settings

Training uses a weighted combination of adversarial, cycle-consistency, and identity losses. We employ the least-squares adversarial objective (LSGAN) implemented with mean squared error (MSE) against binary targets, while cycle-consistency and identity terms are computed using the mean L_1 loss. The total generator objective is:

$$\mathcal{L}_G = \mathcal{L}_{B \rightarrow A}^{adv} + \mathcal{L}_{A \rightarrow B}^{adv} + \lambda_{cyc} \mathcal{L}_{cyc} + \lambda_{id} \mathcal{L}_{id}, \quad (1)$$

where $\lambda_{cyc} = 10$ and $\lambda_{id} = 5$ for all experiments. Here, $\mathcal{L}_{B \rightarrow A}^{adv}$ and $\mathcal{L}_{A \rightarrow B}^{adv}$ denote the Least-Square GAN (LSGAN) generator losses for the two translation directions, $\mathcal{L}_{cyc}$ enforces cycle consistency between the forward and backward mappings, and $\mathcal{L}_{id}$ encourages identity mapping behavior when the input already belongs to the target domain. These weights follow the values established for CycleGAN and are held fixed across every configuration, since varying them alongside the discriminator regularization would confound the two: any change in retrieval could then be attributed to either factor. Fixing them keeps discriminator regularization the only variable under comparison, and the effect of the loss balance is left as an open question noted in the Limitations. Optimization uses Adam with an initial learning rate of 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$ and batch size 1 for 200 epochs. The learning rate is kept constant for the first 100 epochs and then linearly decayed to zero over the remaining 100 epochs using a LambdaLR scheduler. Checkpoints are saved every twenty epochs from epoch 100 onward, giving the candidate set used for validation-based selection in Section 5. The same training protocol is applied consistently across all dataset splits and model configurations.

4. *Discriminator Regularization Strategy*

To stabilize adversarial training and promote identity-preserving synthesis, we regularize both CycleGAN discriminators with spectral normalization and dropout. Spectral normalization is applied to the convolutional layers in all evaluated settings to constrain discriminator sensitivity and improve optimization stability.

The reason for adding dropout is specific to the retrieval task, not to training stability alone. An unregularized discriminator can separate real from synthesized faces using easily exploited pixel-level cues such as local texture, edge sharpness, and tonal distribution. A generator trained against it learns to reproduce those statistics, which need not preserve the identity-discriminative structure an ArcFace embedding depends on, the relative placement of facial landmarks and the proportions of the face. Applying dropout throughout the discriminator forces its decision onto distributed, redundant features, weakening the pixel-level component of the adversarial signal and leaving the generator freer to retain the structural cues retrieval requires.

We therefore apply an all-block dropout strategy, with dropout after the activation in Blocks 1–4 and the output layer left unchanged, acting as a lightweight regularizer that reduces feature co-adaptation and eases overfitting under limited training identities. Dropout was chosen over alternatives such as the R1 gradient penalty or instance noise, which act at the discriminator input or on its input gradients, because our aim is to prevent the discriminator from committing to specific pixel-level cues at every level of its feature hierarchy. We make no claim that dropout is generally superior; the choice follows from the mechanism targeted here. The rates $p = 0.3$ and

$p = 0.5$ (retaining 70% and 50% of activations) lie within the standard dropout range reported by Srivastava *et al.* (2014); the comparison between them is left to the empirical results. The four proposed configurations are R6-Drop30, R9-Drop30, R6-Drop50, and R9-Drop50, denoting ResNet-6 and ResNet-9 generators with all-block dropout at the stated rate.

5. *Retrieval and Evaluation*

Retrieval uses a fixed ArcFace embedding pipeline, the InsightFace buffalo_L model, an ArcFace recognizer (recognition model w600k_r50.onnx, a ResNet-50 backbone) producing 512-dimensional normalized embeddings, run with InsightFace 1.0.1 and ONNX Runtime 1.23.2 at a detection size of 640×640 and a detection threshold of 0.5; when several faces are detected, the one with the largest bounding box is kept. Each synthesized query image $\hat{A}$ passes through face detection, alignment, and embedding extraction to yield a normalized feature vector, and the same pipeline is applied to every gallery photo in the test set. ArcFace serves as a fixed off-the-shelf embedding model without task-specific fine-tuning, so retrieval reflects the usefulness of the synthesized images for downstream recognition rather than adaptation of the recognizer. Identity retrieval then computes cosine similarity between the query embedding and each gallery embedding and ranks gallery identities in descending order of similarity, and the ranked list gives the Top- k performance.

We evaluate from two complementary angles: synthesis quality and retrieval effectiveness. Synthesis quality is assessed in the sketch-to-photo direction ($B \rightarrow A$) by comparing each synthesized photo $\hat{A}$ with the ground-truth test photo A of the same identity, reported as SSIM and PSNR. Retrieval effectiveness uses Top-1, Top-5, and Top-10 accuracy together with the CMC curve, which gives the probability that the correct identity appears within the top- k candidates. All retrieval figures are computed against a closed gallery drawn from the split's 37 test photographs; at this gallery size Top-5 and Top-10 approach their ceiling, so Top-1 carries the comparison and the higher ranks are reported for completeness.

In practice, some synthesized images fail face detection or alignment. An image counts as valid when the detector returns at least one face and a 512-dimensional normalized embedding is extracted from the largest detected bounding box; the same criterion applies to gallery photos. The gallery of each split comprises every test photograph that yields an embedding, assembled independently of whether the corresponding query succeeds, so that all configurations are ranked against an identical gallery (35, 35, 33, 35 and 34 of 37 in splits 1 to 5). Top- k is reported under two denominators: the conditional figure counts only identities whose query and gallery counterpart could both be embedded, while the end-to-end figure takes all 37 test identities and counts a query as incorrect whenever detection, alignment or embedding fails, or whenever its gallery counterpart cannot be embedded. Query embedding success is reported for each setting alongside the retrieval results.

Because the best-performing epoch is not known in advance and cannot be chosen on the test set without leaking information into the evaluation, checkpoints are selected on the validation split. Candidates are saved every twenty epochs from 100 to 200, giving {100, 120, 140, 160, 180, 200}. For each configuration and split, the checkpoint with the highest validation Top-1 is selected, with ties resolved in a fixed order: first by validation Top-5, then by the number of valid validation queries, and finally toward the later epoch. Since the validation gallery holds only 17 to 19 valid queries, validation Top-1 saturates frequently, so this tie-break rule was fixed before any test result was examined. Selection and testing run as separate procedures that never access each other's data: the first reads only the validation split, the second loads the selected checkpoint and evaluates it once on the test split. As the study uses five random identity-disjoint splits, both synthesis and retrieval results are summarized across splits by the mean and sample standard deviation.

Paired Wilcoxon signed-rank tests across the five splits accompany the main comparisons. Because the splits draw on a common identity pool and their test sets overlap, the five paired differences are not independent replicates, so these p -values are descriptive; the direction of the difference on each split and the size of the gap carry the comparison.

Results

1. *Validation-Based Checkpoint Selection*

Because each split draws a different identity-disjoint partition, the epoch at which a model performs best is not fixed in advance, and choosing it on the test set would leak information into the evaluation. Each checkpoint was therefore selected on a held-out validation set, taking the epoch with the highest validation Top-1 among those saved every twenty epochs from 100 to 200, with ties broken in the fixed order specified in Table 2. Table 2 reports the selection for each split and which criterion resolved it.

Validation Top-1 discriminates between candidate epochs far less often than the selection rule implies. Of the thirty selections, six were settled by validation Top-1 alone and one by validation Top-5; the count of valid validation queries never settled a tie, and the remaining twenty-three fell through to the later-epoch rule. Validation Top-1 reaches 1.000 in thirteen of the thirty runs, and ties are wider in the dropout group: eighteen of its twenty selections reached the epoch rule against five of the ten baseline selections, and in three runs all six candidate epochs scored identically. The later selected epochs of the dropout configurations therefore follow in part from the tie-break rule rather than from a measurable difference between epochs.

Table 2 Validation-selected checkpoints for each method under the five-split protocol. Ties in validation Top-1 are broken in order by validation Top-5, the number of valid validation queries, then the later epoch. Selected epochs are listed in split order (1–5); the final column counts selections that reached the last of these criteria.

Method	Selected epochs (splits 1–5)	Resolved by epoch rule
R9-Drop30	200, 200, 200, 180, 200	5 / 5
R6-Drop50	200, 200, 120, 200, 160	5 / 5
R9-Drop50	200, 180, 100, 200, 200	4 / 5
R6-Drop30	200, 200, 180, 120, 200	4 / 5
R6-Base	180, 200, 200, 100, 100	3 / 5
R9-Base	180, 160, 120, 180, 140	2 / 5

Reversing the final tie-break confirms this. Selecting the earliest rather than the latest tied epoch changes twenty-three of the thirty checkpoints and lowers the dropout group's mean selected epoch from 182 to 134, against 156 to 136 for the baselines. Retrieval barely moves: conditional Top-1 shifts by -0.51 points on average, the ordering is identical, and every dropout configuration still exceeds both baselines under either rule. The retrieval finding is therefore independent of the tie-break; the difference in selected epochs is not.

One practical point follows. Selecting on validation rather than the test set means the reported figures are not the peak each method could reach with test-side tuning; they are what it delivers under an honest, identity-disjoint protocol. The no-dropout runs in particular would look stronger if allowed to select on the test set, yet the gap to the dropout group would persist, since their ceiling is lower rather than merely reached sooner.

2. *Cross-Split Retrieval Performance*

The split between dropout and no-dropout runs is the clearest signal in Table 3. Every configuration carrying all-block dropout, R6-Drop30, R9-Drop30, R6-Drop50, and R9-Drop50, lands above both spectral-norm-only baselines on Top-1, and does so on all five splits without exception; end-to-end, the same holds in thirty-nine of the forty configuration-split comparisons, with one tie and no reversal. The two no-dropout runs sit at the bottom, R9-Base at $80.14\% \pm 7.34\%$ and R6-Base at $83.17\% \pm 6.15\%$, while the dropout runs cluster tightly between 91% and 95%, R9-Drop30 reaching the highest mean Top-1 at $94.74\% \pm 2.41\%$.

What matters is the gap between the two families, not the ordering within the dropout cluster. Holding generator depth fixed, adding dropout raises Top-1 by 14.6 points at ResNet-9 (R9-Drop30 over R9-Base) and 8.0

points at ResNet-6 (R6-Drop30 over R6-Base), each winning on all five splits at the minimum two-sided exact $p = 0.0625$ ($W = 0$); end-to-end the same gaps are 13.0 and 6.5 points. The exact p -value does not cross the conventional 0.05 threshold, and because the test sets overlap across splits these p -values are descriptive rather than inferential; the comparison rests on the size of the gaps and their consistency across splits. Depth alone does nothing comparable: R6-Base and R9-Base differ by only 3 points, and neither generalizes well without regularization. The story is dropout, not architecture size.

Table 3 Cross-split retrieval performance at the validation-selected checkpoints (see Table 2), as mean $\pm$ S.D. over five identity-disjoint splits. Conditional figures count only identities embedded on both sides; end-to-end figures use all 37 test identities as the denominator. Best Top-1 per metric in bold.

Method	Query success (%)	Conditional (%)			End-to-End (%)		
		Top-1 (%)	Top-5 (%)	Top-10 (%)	Top-1 (%)	Top-5 (%)	Top-10 (%)
R9-Drop30	94.05 $\pm$ 2.26	94.74 $\pm$ 2.41	99.43 $\pm$ 1.28	100.00 $\pm$ 0.00	87.57 $\pm$ 3.08	91.89 $\pm$ 1.91	92.43 $\pm$ 2.26
R6-Drop50	91.89 $\pm$ 2.70	94.63 $\pm$ 2.54	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	85.95 $\pm$ 2.96	90.81 $\pm$ 1.48	90.81 $\pm$ 1.48
R9-Drop50	91.35 $\pm$ 1.21	94.05 $\pm$ 4.29	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	85.41 $\pm$ 4.10	90.81 $\pm$ 1.48	90.81 $\pm$ 1.48
R6-Drop30	93.51 $\pm$ 2.42	91.16 $\pm$ 4.71	100.00 $\pm$ 0.00	100.00 $\pm$ 0.00	83.78 $\pm$ 5.06	91.89 $\pm$ 1.91	91.89 $\pm$ 1.91
R6-Base	96.22 $\pm$ 2.42	83.17 $\pm$ 6.15	98.23 $\pm$ 1.61	99.39 $\pm$ 1.36	77.30 $\pm$ 5.60	91.35 $\pm$ 3.52	92.43 $\pm$ 3.52
R9-Base	96.22 $\pm$ 2.42	80.14 $\pm$ 7.34	98.82 $\pm$ 1.61	99.39 $\pm$ 1.36	74.59 $\pm$ 8.02	91.89 $\pm$ 3.31	92.43 $\pm$ 3.52

Note: Gallery embedding success was identical across all configurations, at 172 of 185 split-level gallery entries (93.0%).

Within the dropout group the differences are small and follow no consistent pattern with either rate or depth. R9-Drop30 achieves the highest mean Top-1, followed closely by R6-Drop50 and R9-Drop50, and no paired Wilcoxon comparison within the group shows a consistent separation ($p = 0.125 - 1.0$). The earlier claims that moderate dropout outperforms stronger dropout, and that a shallower generator is advantageous, are therefore not supported under the held-out validation-selection protocol.

Top-5 and Top-10 offer little discrimination. With a gallery of this size, differences compress by rank 5, all configurations exceeding 98% and three dropout configurations reaching 100%. R9-Drop30 falls fractionally short only because one query in one split places its correct match outside the Top-5, which given its leading Top-1 is unlikely to signal a ranking weakness. Top-1 thus carries the comparison.

Embedding failures arise on both sides. On the gallery side they are a property of the split, not the model: eight distinct photographs fail, and because some recur across splits these account for 13 of the 185 split-level gallery entries, leaving 172 (93.0%), with the same identities failing wherever they recur. Query failures vary

with configuration, from 1.4 per split for the baselines to 2.2–3.2 for the dropout runs. Most, 57 of 68, fall on identities whose real photograph also fails, pointing to subjects the recognizer handles poorly at this resolution; the rest track with weak facial structure or local artifacts. This runs opposite to the retrieval ordering.

Charging every failure as an error leaves the ordering intact. All four dropout configurations remain above both baselines, R9-Drop30 leading at 87.57% against 74.59% for R9-Base, so the 14.60-point conditional gap at ResNet-9 becomes 12.97 points end-to-end. End-to-end Top-5 and Top-10 sit between 90.81% and 92.43% for every configuration, a ceiling set by embedding success rather than ranking quality, which is why the higher ranks separate the methods even less than under the conditional measure.

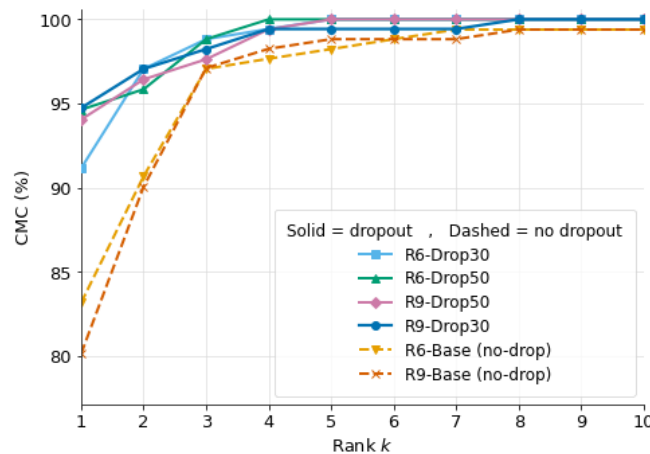


Figure 4 Mean CMC curves across five identity-disjoint splits at the validation-selected checkpoints.

Solid curves denote all-block dropout configurations; dashed curves the no-dropout baselines.

Figure 4 shows the same pattern in curve form. The four dropout curves run together near the top through the early ranks that matter operationally, while the two no-dropout curves trail from rank 1 and close the gap only once everything saturates. The separation is between the solid and dashed curves, a visual restatement that regularizing the discriminator, rather than tuning its rate or the generator's depth, is what moves retrieval.

3. *Cross-Split Synthesis Quality*

Table 4 tells the opposite story to Table 3, and that reversal is the point. The two no-dropout runs, the same configurations that sat at the bottom of the retrieval ranking, produce the highest synthesis fidelity. R9-Base reaches the best SSIM at 0.6930 ± 0.0057 and the best PSNR at 18.23 ± 0.34 dB, with R6-Base close behind. R9-Drop30, the retrieval leader at 94.74% Top-1, sits at 0.6678 SSIM and 16.63 dB, in the lower half of the fidelity table.

Table 4 Cross-split synthesis quality (SSIM, PSNR) at the validation-selected checkpoints (see Table 2), reported as mean $\pm$ S.D. across five identity-disjoint splits. Methods are ordered as in Table 3. Highest fidelity per column in bold; each falls on a no-dropout run, unlike the best Top-1 in Table 3.

Method	Mean SSIM	Mean PSNR (dB)
R9-Drop30	0.6678 $\pm$ 0.0132	16.63 $\pm$ 0.78
R6-Drop50	0.6417 $\pm$ 0.0099	16.80 $\pm$ 0.42
R9-Drop50	0.6417 $\pm$ 0.0101	16.78 $\pm$ 0.43
R6-Drop30	0.6718 $\pm$ 0.0047	17.04 $\pm$ 0.44
R6-Base	0.6840 $\pm$ 0.0062	17.79 $\pm$ 0.61
R9-Base	0.6930 $\pm$ 0.0057	18.23 $\pm$ 0.34

This inversion is direct evidence that SSIM and PSNR do not predict retrieval in this task: a method can reconstruct a photo closer to the ground truth in pixel and structural terms yet fail more often at placing the correct identity at rank 1. The no-dropout runs give the smoother, higher-fidelity reconstructions, but their embeddings separate identities less cleanly once passed through ArcFace. The point holds within the dropout group, where SSIM ranges only from 0.6417 to 0.6718 and bears no relation to retrieval: R6-Drop30 has the highest SSIM of the four yet the lowest Top-1.

Pixel-faithful and identity-preserving synthesis are therefore different objectives, and selecting for the former can work against the latter. Dropout in the discriminator trades a little reconstruction fidelity for embeddings that retain identity-discriminative structure, a trade fidelity metrics penalize but retrieval rewards. For a forensic pipeline, where the deliverable is a correct ranked identity rather than a faithful portrait, it is the trade worth making, and it is why SSIM and PSNR should not serve as selection or stopping criteria here.

4. *Qualitative and Quantitative Comparison of Synthesized Photos*

The numbers in Tables 3 and 4 establish that fidelity and retrieval pull apart; the synthesized images show what that separation looks like. Figure 5 presents six identities from the representative split, arranged so the visual differences between methods can be read against the difficulty of each case. The six were chosen without reference to any model output: within the split whose retrieval sits closest to the five-split mean, every test identity was scored by the cosine similarity between its real sketch and real photo embeddings, and two identities were taken from the high, median, and low ends of that ordering. These three bands form the easy, medium, and hard

rows of the figure. Each row runs from the input sketch through the two no-dropout baselines and the four dropout variants to the ground-truth photo.

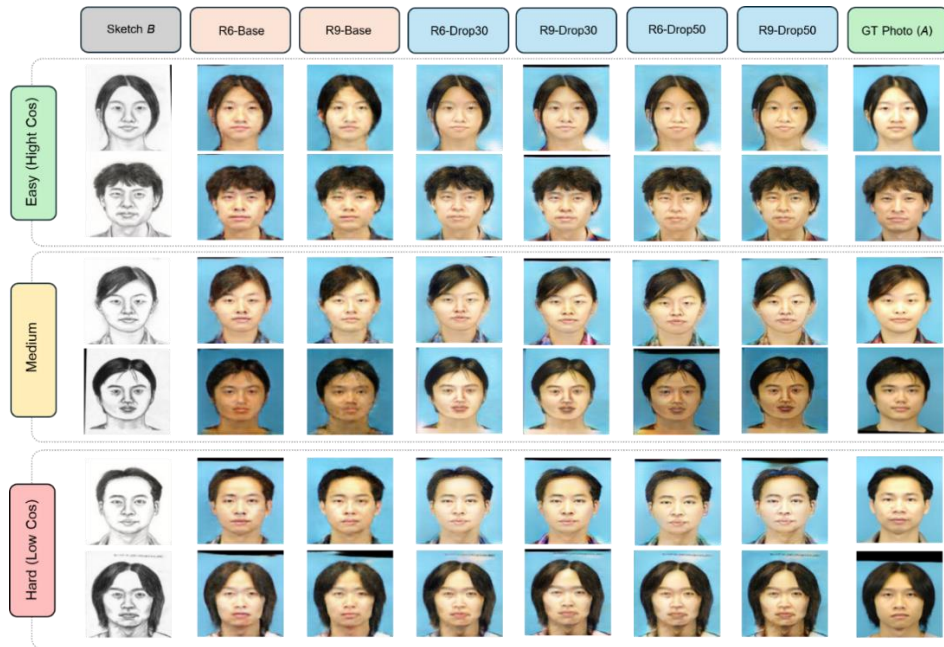


Figure 5 Qualitative comparison of synthesized photos for six identities from the representative split (retrieval closest to the five-split mean). Identities were ranked by cosine similarity between their real sketch and real photo embeddings, sampling two from the high (easy), median (medium), and low (hard) ends, independently of any model output. Each row shows, left to right: input sketch (Domain B), two no-dropout baselines, four dropout variants, and ground-truth photo (Domain A). Checkpoints follow Table 2.

For the easy identities, the methods are hard to tell apart. All six reconstructions recover the facial layout and read as the same person as the ground truth, with only R9-Drop30 looking slightly sharper. When the sketch already carries strong identity information, the choice of regularization makes little visible difference. The methods separate as the cases grow harder, and the divide falls along the dropout boundary. The two no-dropout baselines render skin as a smooth, evenly toned surface, which looks clean but sacrifices fine structure. Hair shows this most clearly: in both medium identities the no-dropout outputs blur the strands into a flat mass, and in the hard row the long-haired male subject loses the center parting entirely, a cue the dropout variants keep. The dropout outputs carry a rougher surface texture yet hold onto the small, high-frequency details that tell one face from another. One hard identity has a mole beside the nostril in both sketch and photo; only R9-Drop30 reproduces it, faint but present. The dropout variants also add light and shading that give the harder faces more depth, most visibly in the second hard

identity. This pattern is the visual counterpart of the numbers. Smoothing the skin is what lifts the pixel-level fidelity of the no-dropout runs, and the same smoothing erases the local cues an embedding relies on, which is why the method with the highest SSIM in Table 4 is not the one that retrieves best. Pushing the regularization harder does not simply extend the benefit: at $p = 0.5$ the outputs begin to show faint double edges, a sign of over-regularization, consistent with the earlier finding that having dropout matters more than setting it to its strongest value.

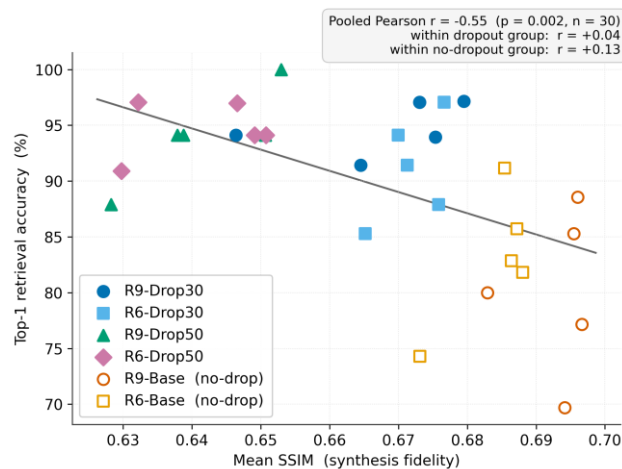


Figure 6 Synthesis fidelity (mean SSIM) versus Top-1 retrieval across all thirty method-split combinations (six configurations $\times$ five splits). Filled markers denote dropout configurations; open markers the no-dropout baselines. The two are negatively correlated (Pearson $r = -0.55$, $p = 0.002$): baselines achieve the highest SSIM yet the lowest retrieval, while within either family the correlation is negligible.

Figure 6 makes the relationship explicit. Each point is one configuration on one split, its mean SSIM against its Top-1 retrieval, and the thirty points slope downward rather than up: fidelity and retrieval are negatively correlated (Pearson $r = -0.55$, $p = 0.002$). The two no-dropout baselines occupy the high-fidelity, low-retrieval corner, exactly where the qualitative comparison placed them, while the dropout variants trade a little SSIM for a marked gain in retrieval. The trend is a between-family effect; within either family the correlation is negligible (Pearson $r = +0.04$ for the dropout configurations, $+0.13$ for the baselines), so it reflects the presence or absence of regularization rather than any tuning of the drop rate. The same decoupling holds for individual images: across the 1,021 conditional queries pooled over all splits and configurations, per-image SSIM and retrieval rank are effectively uncorrelated (Spearman $\rho = -0.04$). Pixel-level fidelity therefore did not provide a reliable basis for selecting or ranking models for identity retrieval here; a model chosen for its SSIM would be among the weakest retrievers in this study.

Discussion

The results converge on a single mechanism, and three questions organize the discussion: why discriminator dropout raises retrieval while lowering reconstruction fidelity, why generator depth does not decide the outcome when regularization does, and what this means for a forensic retrieval system built from synthesized faces. The training dynamics in Figure 7, shown below, anchor all three answers.

The training curves bear on the first question. Figure 7(c) shows the validation generator loss splitting cleanly along the dropout boundary: both baselines fall to a shallow minimum near epoch 20, then rise for the rest of training, reaching about 14.3 by epoch 200 despite the learning-rate decay that begins at epoch 100, while the four dropout runs hold between 1.57 and 1.68 throughout. A validation loss that grows while training continues is consistent with a discriminator overfitting its own task, its adversarial signal no longer generalizing as it separates real from synthesized faces more confidently. This is consistent with their weaker retrieval in Table 3.

The training-side panels show the same imbalance and confirm optimization stays stable. In Figure 7(b) the baselines' discriminator loss collapses toward zero (near 0.02) while the dropout runs settle around 0.25, a gap present from the first epoch; Figure 7(a) shows every generator loss descending smoothly to a plateau. Dropout keeps the adversarial game balanced rather than making synthesis easier, and it is that balance that preserves the identity-discriminative structure ArcFace relies on: a discriminator forced onto distributed, redundant features presents a weaker pixel-level objective, leaving the generator freer to retain the structural cues that separate faces. Those cues survive at some cost to smoothness, which is why the highest-SSIM model retrieves worst.

Depth tells a quieter story, and that contrast is the point. With regularization fixed, moving from ResNet-6 to ResNet-9 barely shifts retrieval (baselines differ by three points, $p = 0.375$, no consistent direction), while adding dropout at fixed depth lifts Top-1 substantially and the same way on every split ($p = 0.0625$, $W = 0$). Extra capacity helps only when the signal driving it is well behaved. The rate is likewise non-decisive: $p = 0.3$ and $p = 0.5$ show no consistent difference despite a higher discriminator loss at the stronger rate. What governs retrieval is the presence of regularization, not its strength or the generator's size.

The implication for forensic use follows from the same evidence. The intuitive criterion, the most faithful reconstruction, points the wrong way: the best-SSIM model retrieves worst, and selecting on fidelity would have chosen it. Since the deliverable is a correct ranked identity, not a faithful portrait, checkpoints and stopping points should be judged on retrieval against a held-out gallery, as our validation-Top-1 protocol does, rather than on fidelity or loss.

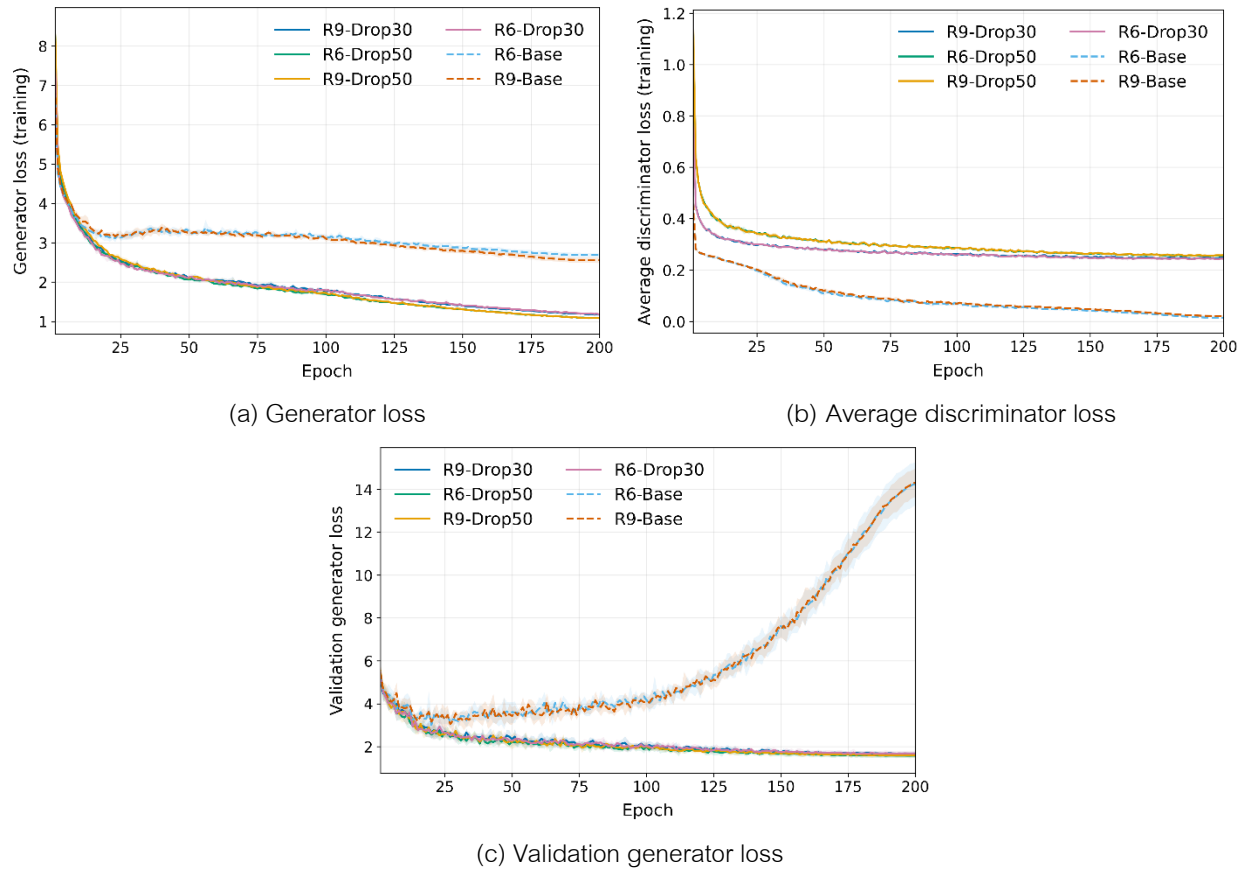


Figure 7 Training dynamics across the five identity-disjoint splits, shown as mean (line) $\pm$ S.D. (shaded band) per epoch. (a) Generator loss and (b) average discriminator loss (mean of D_A and D_B), both on the training set; (c) generator loss on the held-out validation set. Panel (c) shows the divergence central to the discussion: baselines fall to a shallow minimum before rising, while dropout runs stay low and flat through all 200 epochs.

Conclusions

This work set out to build a sketch-to-photo pipeline whose synthesized faces are useful for identity retrieval rather than merely pleasant to look at, and to test what actually drives that usefulness. Working from a CycleGAN backbone with a spectral-normalized discriminator, we added all-block dropout to the discriminator and evaluated every configuration through a fixed ArcFace recognizer and cosine retrieval, across five identity-disjoint splits of CUFS with checkpoints chosen on a held-out validation set. Pairing each dropout model with a no-dropout baseline at the same generator depth, and adding the ResNet-6 baseline that earlier work lacked, lets the comparison change one variable at a time.

Three findings come out of that setup and together form the contribution. The first is the pipeline itself, a retrieval-oriented sketch-to-photo system that treats the correct ranked identity, not the faithful portrait, as the quantity to optimize, evaluated as a complete pipeline under a validation-based protocol built to avoid leakage. The second is the discriminator regularization strategy: all-block dropout on top of spectral normalization lifts Top-1 retrieval consistently over the unregularized baselines, with no split favoring a baseline under either denominator, while the isolating baseline shows that generator depth alone does not account for the gain. What matters is the presence of the regularization, not its exact rate or the size of the generator. The third, and the one with the widest reach, is that pixel fidelity does not predict retrieval in this task: the no-dropout baselines produce the highest SSIM and PSNR yet retrieve worst, the two correlate negatively across configurations, and the training curves are consistent with a discriminator that overfits without regularization and stays balanced with it.

The practical message for forensic use follows from the third finding. A face that scores well on reconstruction metrics is not thereby a better search key, so a system built to return the right identity should select and stop its synthesis model on retrieval against a held-out gallery, not on SSIM, PSNR, or a low training loss. The multi-split, validation-based protocol used here is best seen as the evaluation practice that makes such a claim trustworthy rather than as a contribution in its own right.

Limitations and Future Work

This study is limited by the CUFS dataset, which contains relatively few identities and clean, standardized sketches that may not reflect the noise and stylistic variability of real forensic sketches. The closed-set protocol assumes the true identity is always present in the gallery, an assumption that may not hold in an operational search, and with 37 test identities the higher ranks saturate, leaving Top-1 as the only informative measure. Five paired splits place a floor of 0.0625 on the attainable Wilcoxon p -value, and because the test sets overlap the paired differences are not independent replicates; we therefore treat these p -values as descriptive. A validation gallery of 17 to 19 valid queries likewise limits the resolution of checkpoint selection, leaving most ties to the fixed rule. Loss weights and the ArcFace recognizer were held fixed to isolate discriminator regularization, and direct comparison with external methods is limited because prior studies differ in synthesis direction, recognizer, dataset, and evaluation protocol. Future work should evaluate the approach against a diffusion-based baseline on an identical retrieval protocol, extend it to larger and noisier datasets and to open-set retrieval, and test whether the retrieval gains persist across face-recognition backbones and under end-to-end fine-tuning. Beyond testing the framework on new data or backbones, a further direction follows from the finding that fidelity does not predict retrieval:

incorporating the retrieval objective into training itself, through a retrieval-aware loss or stopping criterion that optimizes identity ranking directly.

Data Availability

The identity split manifests (train/validation/test lists for all five seeds), model configurations, and evaluation protocol are available at <https://anonymous.4open.science/r/sketch-photo-retrieval-dropout-1D8B/>. The CUFS dataset is the property of its original authors and is not redistributed; it can be obtained from its official source and combined with the provided manifests to reproduce the splits.

Acknowledgments

The authors gratefully acknowledge the Navaminda Kasatriyadhiraj Royal Air Force Academy for providing institutional support and the computing environment used in this study. The authors also acknowledge the CUFS dataset, which served as the primary benchmark for the sketch-photo synthesis and retrieval experiments reported in this paper.

References

- Chugh, T., Singh, M., Nagpal, S., Singh, R., & Vatsa, M. (2017). Transfer learning based evolutionary algorithm for composite face sketch recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1117–1125).
- Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4690–4699.
- Du, K., Wang, C., Cao, L., Guo, Y., & Sun, W. (2024). A modality separation approach for facial sketch synthesis. *IET Image Processing*, 18, 4127–4140. <https://doi.org/10.1049/ipr2.13238>
- Fang, Y., Hu, J., & Deng, W. (2020). Identity-aware CycleGAN for face photo-sketch synthesis and recognition. *Pattern Recognition*, 102, 107249. <https://doi.org/10.1016/j.patcog.2020.107249>
- Guo, Y., Cao, L., Chen, C., Du, K., & Fu, C. (2020). Domain alignment embedding network for sketch face recognition. *IEEE Access*, 9, 872–882. <https://doi.org/10.1109/ACCESS.2020.3047108>

- Huang, Z., Tang, F., Zhang, Y., Cao, J., Li, C., Tang, S., Li, J., & Lee, T.-Y. (2024). Identity-preserving face swapping via dual surrogate generative models. *ACM Transactions on Graphics*, 43(5), Article 161, 1–19. <https://doi.org/10.1145/3676165>
- Mahfoud, S., Daamouche, A., Bengherabi, M., & Hadid, A. (2022). Hand-drawn face sketch recognition using rank-level fusion of image quality assessment metrics. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 70(6), e143554. <https://doi.org/10.24425/bpasts.2022.143554>
- Mahfoud, S., Bengherabi, M., Daamouche, A., Boutellaa, E., & Hadid, A. (2024). High-quality synthesized face sketch using generative reference prior. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 72(4), Article 150109.
- Phisutthangkoon, N., Anchuen, P., Luangwilai, T., & Daungklang, P. (2026). Face photo and sketch synthesis using CycleGAN: The role of residual blocks and spectral normalization. *International Scientific Journal of Engineering and Technology*, 10(1), 84–95.
- Que, Y., Xiong, L., Wan, W., Xia, X., & Liu, Z. (2024). Denoising Diffusion Probabilistic Model for Face Sketch-to-Photo Synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10), 10424–10436. <https://doi.org/10.1109/TCSVT.2024.3409184>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958.
- Tang, D., Jiang, X., Wang, K., Guo, W., Zhang, J., Lin, Y., & Pu, H. (2024). Toward identity preserving in face sketch-photo synthesis using a hybrid CNN-Mamba framework. *Scientific Reports*, 14, 22495. <https://doi.org/10.1038/s41598-024-72066-y>
- Tang, D., Jiang, X., Zhang, Y., Dai, Y., & Lin, Y. (2025). Ipdpm: identity preserving diffusion model for face sketch and photo synthesis. *Machine Vision and Applications*, 36, Article 34. <https://doi.org/10.1007/s00138-024-01658-5>
- Tang, X., & Wang, X. (2003). Face sketch synthesis and recognition. In *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003)* (Vol. 2, pp. 687–694).

- Wan, W., Yang, Y., & Lee, H. J. (2021). Generative adversarial learning for detail-preserving face sketch synthesis. *Neurocomputing*, 438, 107–121. <https://doi.org/10.1016/j.neucom.2021.01.050>
- Wang, N., Zhu, M., Li, J., Song, B., & Li, Z. (2017). Data-driven vs. model-driven: Fast face sketch synthesis. *Neurocomputing*, 257, 214–221. <https://doi.org/10.1016/j.neucom.2016.07.071>
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.