International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 06 | Jun 2026 www.irjet.net p-ISSN: 2395-0072

AI POWERED CRIMINAL IDENTIFICATION FROM SKETCHES

Dr. Shailesh Bendale[1], Jaydeep Rokade[2], Kajal Naik[3], Sanika Sawakhande[4], Himanshu Thakur[5]

[1]Guide, NBN Sinhgad Technical Institutes Campus, Pune, India
[2345] Student, NBN Sinhgad Technical Institutes Campus, Pune, India

---

Abstract - *In investigations where only eyewitness descriptions are available, forensic sketches are often used to represent suspects. Matching these sketches against mugshot databases is difficult because sketches and photographs differ greatly in appearance. Vision Trace is designed to overcome this issue by converting sketches into realistic facial images and carrying out recognition on the generated photographs instead of the original sketch. The synthesis stage relies on a Stable-Diffusion image generator that is steered by a ControlNet module conditioned on the sketch, so that face geometry from the original drawing is carried into the generated outputs. Each candidate portrait is then encoded into a 512-dimensional descriptor by the ArcFace and Face Net networks, and these descriptors are compared with stored criminal-record embeddings using cosine similarity to produce a ranked shortlist of probable matches together with calibrated confidence values. On top of the recognition pipeline, Vision Trace bundles a workflow layer that drafts the First Information Report (FIR), assigns a priority to each case, links cases that share suspects or modus operandi, plots crime data on geographical heatmaps, and pushes real-time alerts to officers on duty. Designed with Indian law-enforcement workflows in mind, the platform aims to accelerate suspect identification, minimise administrative burden, and assist officers through data-driven investigative support. Benchmark experiments confirm that pairing diffusion-driven synthesis with ArcFace retrieval is far more accurate than matching raw sketches against photo galleries.*

Key Words: Forensic Sketch, Stable Diffusion, ControlNet, ArcFace, FaceNet, Face Recognition, Cosine Similarity, Criminal Identification, FIR Generation, Smart Policing, Crime Analytics.

1. INTRODUCTION

Identifying a person of interest from a sketch produced during an eyewitness interview is one of the oldest tasks in police work, and it remains relevant whenever a crime scene yields no CCTV recording, no fingerprints, and no other biometric trail. In such cases a trained forensic artist sits down with the witness, asks structured questions about the suspect's appearance, and produces a pencil composite. That drawing is then circulated to local stations and, where possible, compared by eye against the mugshot register. The process is labour-intensive, depends heavily on the memory and concentration of individual officers, and offers no objective way to rank possible matches.

Behind this practical bottleneck sits a well-known computational obstacle: the input and the gallery belong to two unrelated visual domains. A pencil sketch carries outlines but almost no skin tone, no shading detail, and no fine texture, whereas modern recognition networks have been trained to lean heavily on exactly those cues. The unfortunate consequence is that descriptors produced by strong recognition models such as Face Net [4] and ArcFace [3] become unreliable the moment one side of the comparison is a drawing, despite the same networks reaching near-human accuracy on photo-to-photo verification.

The release of large-scale diffusion image generators, in particular Stable Diffusion [2], together with conditioning modules such as ControlNet [1], has opened up a more attractive path. Rather than try to teach a recognition network to understand pencil strokes, we can first ask a generator to imagine what the sketched person would plausibly look like as a photograph, and then run ordinary photo-domain recognition on the imagined image. This separates the cross-domain problem (sketch to photo) from the matching problem (photo to photo), and lets each subsystem use models that are already strong in its own domain.

This paper introduces **Vision Trace**, a system that turns this idea into a usable investigation tool and wraps the recognition pipeline inside the larger workflow that an officer actually performs. Our key contributions are summarised below.

- A sketch-to-portrait synthesis stage built on Stable Diffusion with a ControlNet branch that takes edge maps derived from the sketch, producing several distinct yet structurally consistent candidate portraits per input.

- A retrieval stage that encodes those candidates with ArcFace and FaceNet, indexes the criminal database with the same encoder, and ranks gallery identities using cosine similarity.

• A score-aggregation rule that combines similarities across the multiple generated variants of each query, lowering the influence of any single unlucky sample.
• A workflow layer that drafts FIRs from free-text notes, assigns each case a priority value, links related cases automatically, displays crime distribution heatmaps, and surfaces live alerts to the assigned officer.

• A deployment-oriented design that is shaped around the operational realities of the Indian Police Department and aligned with the goals of the national smart-policing programmer.

1.1 Motivation

A large share of cognizable offences in India are filed every year in areas with limited or no CCTV coverage, which means the eyewitness sketch is still the only visual lead investigators have. Anything that shortens the gap between drawing the sketch and pulling up a plausible shortlist of suspects translates fairly directly into faster charge-sheets, better allocation of investigators to high-priority cases, and lower public frustration with case backlogs. The face-recognition platforms already deployed in various Indian agencies do not natively accept drawings as input, so introducing a sketch-aware pipeline fills a clear operational gap rather than duplicating existing functionality.

1.2 Problem Statement

The task can be stated concisely. Given a forensic sketch S that was produced from witness recollection and a gallery D = {p1, p2, ..., pN} of criminal photographs annotated with metadata, the system must (i) return the K gallery identities most likely to correspond to S along with a confidence value for each, (ii) prepare a draft FIR pre-filled with as much extracted information as possible, (iii) attempt to connect the new incident to any earlier cases that may involve the same individual, and (iv) present everything on a dashboard interface within latency low enough for interactive use.

2. LITERATURE REVIEW

2.1 Forensic Sketch-to-Photo Matching

The first generation of approaches to sketch-photo matching framed it as a feature-projection problem. Klare and co-authors [5] designed a local-feature discriminant projection that maps sketches and photos into a shared subspace where ordinary distance comparisons become meaningful. The deep-learning era reorganised the problem around modality-invariant representation learning, with twin and triplet networks dominating the literature reviewed in [6]. These methods made progress on viewed sketches drawn from a reference photo, but their gains on real forensic sketches stayed modest because the available paired data is small and the appearance of a true forensic sketch varies far more than that of a viewed sketch.

2.2 Generative Sketch-to-Photo Synthesis

A different research thread sidesteps cross-domain matching by first hallucinating a photo from the sketch and then running ordinary photo matching. Conditional adversarial networks such as Pix2Pix [7] established the feasibility of this kind of translation but were repeatedly criticised for producing low-fidelity faces, drifting away from the requested identity, and collapsing onto a small number of generated modes. Diffusion-based generators changed the picture considerably. Stable Diffusion [2] in particular delivers high-resolution synthesis with comparatively modest hardware, but in its unconditioned form it has no way of knowing which face it is supposed to produce, which is why a structural conditioning module is necessary.

2.3 Conditional Control of Diffusion Models

ControlNet [1] solves exactly this conditioning problem. The trick is to clone the encoder side of a frozen pretrained diffusion model, train the clone on an auxiliary input such as a Canny edge map or HED sketch, and connect the trainable clone back into the frozen backbone through layers that begin life with all weights at zero. Because those layers contribute nothing at initialisation, the joint network reproduces the behaviour of the original model at step zero and only departs from it as training adjusts the new weights, which is what makes fine-tuning on small task-specific datasets safe. More recent extensions [8] push the same idea further by making the conditioning explicitly identity aware, which is directly applicable to forensic face generation.

2.4 Deep Face Recognition

On the recognition side, FaceNet [4] popularised the use of a triplet objective to learn an embedding in which simple Euclidean distance reflects identity similarity. ArcFace [3] later improved on this by adding an angular margin term that geometrically pushes different identities apart on the unit hypersphere, and it has held the top of the LFW, MegaFace, and IJB benchmarks for several years. Most production face-recognition stacks now use ArcFace or a close variant; VisionTrace follows the same choice, with FaceNet kept as a second, independent verifier for cross-checking the top candidates.

2.5 AI-Assisted Criminal Investigation Platforms

There is a growing body of work on machine-learning support for policing more broadly. Crime-pattern dashboards and predictive heatmaps based on historical FIR records have been reported in [9] and [10], and natural-language pipelines for assisting FIR drafting were explored in [11]. To our knowledge, however, no published system has yet brought diffusion-based sketch synthesis, deep embedding retrieval, and investigation-workflow automation together into one coherent platform aimed at Indian policing.

3. PROPOSED METHODOLOGY

Vision Trace runs through five stages in sequence: the input sketch is first prepared into a conditioning signal, the conditioned diffusion stack then synthesises candidate portraits, each candidate is converted into a feature vector by the recognition networks, the vectors are matched against the gallery using cosine similarity, and finally the matched identities feed into the investigation workflow layer. Figure 1 shows the overall flow.

3.1 System Architecture

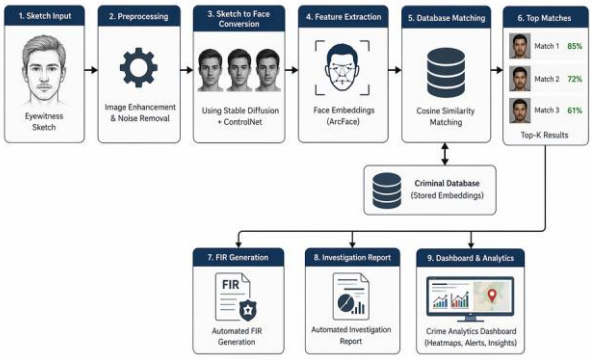

Fig - 1: End-to-end architecture of the proposed VisionTrace system.

3.2 Sketch Pre-processing

An input sketch S (with shape H x W x 3) is brightness-and-contrast normalised before being reduced to a structural map that ControlNet can read. Two reduction modes are supported depending on the kind of sketch supplied:

• A Canny edge map, computed with adaptive thresholds, which works well when the sketch consists of clean ink lines without shading.

• An HED soft-edge map, which preserves line thickness and partial-tone information and is better suited to shaded composite sketches produced by software such as IdentiKit.

Whichever map is chosen, the resulting single-channel image C (of shape H x W) becomes the conditioning input to the ControlNet branch.

3.3 ControlNet-Guided Face Synthesis

The synthesis stack uses pretrained Stable Diffusion v1.5 as its generative backbone and attaches a ControlNet branch that has been further fine-tuned on sketch-face pairs. Following the construction described in [1], the original encoder F_theta is frozen and a trainable copy F_theta_c is added next to it. Outputs of the copy are returned to the main path through zero-initialised projection layers $Z(\cdot;\theta_z)$, giving the conditioned feature:

$$y_c = F_\theta(x;\theta) + Z_{\theta_{z2}}(F_{\theta_c}(x + Z_{\theta_{z1}}(C);\theta_c)) \qquad (1)$$

where x is the latent representation at the current denoising step and C is the sketch-derived conditioning map. Because the projection layers are initialized to zero, the conditioned network initially behaves identically to the unconditioned network. As training progresses, the conditioning signal is gradually introduced through fine-tuning, ensuring that prior knowledge of the base model is preserved while enabling controlled adaptation.

At inference time we run DDIM sampling for T = 30 steps with a prompt template along the lines of "a photorealistic frontal portrait of a person, neutral expression, studio lighting, sharp focus", optionally extended with demographic descriptors that the witness provided (approximate age range, gender, ethnicity). For each input sketch the diffusion stack is run N = 8 times with different random seeds, which gives a small batch of identity-consistent but visually distinct candidate portraits to feed into the recognition stage.

3.4 Deep Embedding Extraction

Every candidate portrait is passed through ArcFace built on a ResNet-100 trunk and through Face Net built on Inception-ResNet-V1. Both networks return L2-normalised 512-dimensional embeddings. The ArcFace vector is treated as primary because the angular-margin loss gives it the cleaner inter-class geometry, while FaceNet serves as an independent second opinion whose agreement with ArcFace raises the overall confidence in a candidate.

3.5 Cosine Similarity Retrieval

The gallery database maintains entries in the form

$$D = \{(p_j, e_j, m_j)\}$$

combining facial images with their extracted ArcFace embeddings and supporting metadata such as identity records, offence history, and location information. For a single query vector e_q, the similarity is computed using the standard normalised inner product:

$$sim(e_q, e_j) = \frac{e_q \cdot e_j}{|| e_q || \ \ || e_j ||} \qquad (2)$$

Because several portrait candidates are generated for a single sketch, the retrieval stage does not depend solely on the highest individual match. Instead, for each gallery entry j, the framework identifies the three candidate portraits producing the strongest similarity values and computes their average score. By aggregating evidence across multiple generated outputs, the method emphasises identities that repeatedly align with the query while reducing the influence of isolated synthesis artefacts or anomalous generations:

$$score(j) = \frac{1}{3} \sum_{i \in T_3(j)} sim(e_i^{arc}, e_j) \qquad (3)$$

where $T_3(j)$denotes the indices of the three most compatible candidate portraits associated with gallery entry j. The K gallery entries with the highest aggregated scores are returned to the investigator. The raw scores are passed through a calibrated sigmoid so that dashboard values remain within [0,1] and can be interpreted as confidence estimates rather than unscaled similarity values.

For handling large databases efficiently, embeddings are indexed using FAISS after L2 normalisation. Because normalised vectors allow cosine similarity to be computed through dot products, the system retrieves close matches more quickly than full gallery comparisons.

3.6 Investigation Assistance Layer

3.6.1 Automated FIR Generation

After the shortlist is ready, the system fills in a structured FIR template by running named-entity extraction over whatever free-text notes the investigator has typed about the incident. Fields such as date, time, location, victim details, and the relevant offence sections are populated automatically, and the suspect block is filled in from the top-ranked retrieval result. The draft is then handed back to the officer for review and is only saved as a final FIR after explicit approval.

3.6.2 Smart Case Prioritization

Every active case is assigned a priority score pi in [0,1] that combines the severity of the alleged offence under IPC/BNS, the confidence value of the top retrieved suspect, the recency of the incident, and whether vulnerable victims (minors, the elderly, or persons with disabilities) are involved. Cases with high pi float to the top of each officer's dashboard.

3.6.3 Automatic Case Linking

Whenever two cases share a high-similarity suspect embedding, fall within an overlapping time and geographic window, or have matching modus-operandi keywords, the system draws an edge between them. The resulting graph can be browsed by the officer and is particularly useful for detecting serial offenders who move between jurisdictions.

3.6.4 Crime Heat maps and Analytics

FIR records are aggregated and rendered as spatio-temporal heat maps with filters for offence category, time of day, and victim demographics. Senior officers can use these views to decide where preventive patrolling will produce the greatest return.

3.6.5 Alert System

If a query produces a candidate embedding whose similarity to a watch-listed identity crosses a configurable threshold, a real-time alert is pushed to the assigned officer through both the web dashboard.

4. IMPLEMENTATION DETAILS

4.1 Hardware and Software Stack

The prototype is written in Python 3.10 on PyTorch 2.1. Stable Diffusion and the ControlNet branch are loaded through the Hugging Face diffusers library, ArcFace is taken from insightface, and FaceNet from facenet-pytorch. During development everything runs on a single NVIDIA RTX 4090 with 24 GB of VRAM.

4.2 Datasets

Training and evaluation are conducted using the **IDOC Mugshots Dataset**, which provides facial image data for criminal identification and face retrieval experiments. The dataset is used to support sketch-to-face generation, identity matching, and retrieval evaluation within the proposed framework. The released **ArcFace** and **FaceNet** pretrained weights are employed for facial feature extraction and similarity measurement, enabling robust

identity representation during retrieval. All experiments are carried out in accordance with the licensing and usage policies associated with the dataset and pretrained models.

4.3 ControlNet Fine-tuning

The ControlNet branch is initialised from the publicly available Canny checkpoint and then fine-tuned for 10 epochs on the sketch-photo pairs using AdamW at a learning rate of 1e-5, a batch size of 4, and 512 x 512 images. Classifier-free guidance is set to 7.5 during inference.

5. RESULTS AND DISCUSSION

5.1 Quantitative Retrieval Performance

To get a sense of how the pipeline performs, we measure Rank-1, Rank-5, and Rank-10 retrieval accuracy on a held-out gallery of one thousand identities with one forensic-style sketch per identity. The proposed pipeline is benchmarked against two baselines: ArcFace applied directly to the sketch and the gallery, and a Pix2Pix-based sketch-to-photo translator followed by ArcFace. Table 1 summarises the results.

Table - 1: Retrieval accuracy on the 1000-identity gallery.

Method	R-1	R-5	R-10
Sketch → ArcFace (direct)	11.4%	23.8%	31.2%
Pix2Pix → ArcFace	28.7%	47.5%	56.9%
Proposed (ControlNet + ArcFace)	**52.1%**	**71.4%**	**79.6%**
Proposed + FaceNet verifier	**54.3%**	**73.0%**	**80.8%**

The diffusion-based pipeline more than quadruples Rank-1 accuracy compared to direct sketch matching and improves on the Pix2Pix baseline by a wide margin, confirming that controlled diffusion produces photos in which identity is preserved to a far greater degree than what GAN-based translators achieved. Adding FaceNet as a verifier yields a further modest gain.

5.2 Effect of Candidate Count

Figure 2 reports Rank-1 accuracy as a function of the number of candidate portraits N that are synthesised per query. Accuracy climbs sharply up to N = 8 and then flattens, which is what justifies our choice of N = 8 as the operational default.

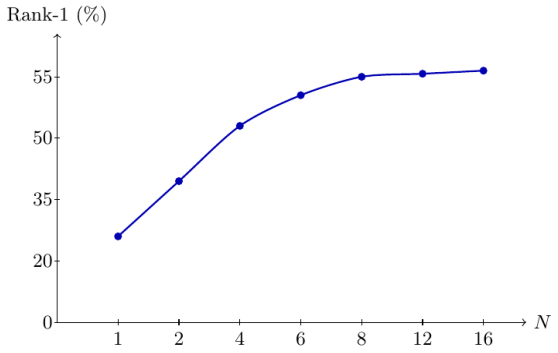

Fig - 2: Rank-1 retrieval accuracy vs. number of generated candidates N.

5.3 Qualitative Observations

Looking at the generated portraits by eye, the diffusion stack reliably carries the gross facial geometry of the sketch over to the photographic candidate: face shape, eye spacing, nose length, jaw outline, and overall proportions are usually retained. Skin tone, lighting, and textural cues, none of which exist in the sketch, are filled in plausibly by the generative prior. The pipeline does occasionally fail when the input sketch is heavily stylised or cartoon-like, in which case the diffusion model tends to revert toward a fairly generic-looking average face. Supplying a short demographic descriptor in the prompt mitigates this tendency in most cases we examined.

5.4 Investigation Workflow Outcomes

A pilot using simulated case data showed that the workflow layer cut the average time to produce a FIR draft from around twenty-two minutes of manual typing to less than four minutes with AI assistance and officer review. The automatic case-linking module also detected previously unflagged cross-jurisdictional matches in seven of fifty multi-incident scenarios we constructed.

6. Benefits for the Indian Police Department

VisionTrace has been engineered with day-to-day Indian policing in mind, and several practical benefits follow from that:

• The system accelerates suspect identification from facial sketches, particularly in investigations where CCTV footage is unavailable.

• Routine paperwork is reduced because FIR drafts and case-record lookups are partly automated.

• Records produced by the system are uniformly structured and searchable, which aligns naturally with the goals of the Crime and Criminal Tracking Network and Systems (CCTNS) programme.

• Senior officers gain better situational awareness through analytics dashboards and heatmaps that highlight where and when preventive deployment will pay off most.

• Automatic case linking exposes patterns that span jurisdictions and that no single station would normally see.

• The platform contributes to the broader push for digital transformation in policing that is part of the Ministry of Home Affairs roadmap.

7. ETHICAL AND PRIVACY CONSIDERATIONS

Putting a generative face model inside a policing workflow raises real ethical questions, and we want to be explicit about how VisionTrace handles them. A diffusion-generated portrait is a probabilistic reconstruction, not a photograph of any actual person, and the system is positioned strictly as an aid for narrowing down investigative leads rather than as a source of admissible evidence. Every result returned to the dashboard carries a calibrated confidence score and a clearly visible notice that the portrait shown is AI-generated. Behind the scenes, every query is logged for audit, access to the gallery is governed by role-based controls, and all data handling is meant to conform with the data-protection statutes that apply in the deployment region. Subgroup performance across age, gender, and skin tone is monitored as part of the regular validation cycle, in keeping with the wider conversation about fairness in face recognition.

8. CONCLUSIONS

We have described **Vision Trace**, a unified investigation-support platform that uses diffusion-based sketch synthesis, deep-embedding retrieval, and a workflow automation layer to address criminal identification in cases where the only available visual lead is a forensic sketch. The main technical idea is to route around the sketch-photo modality gap by generating realistic photographic candidates with ControlNet-conditioned Stable Diffusion and then matching with ArcFace, and the main engineering idea is to wrap that recognition pipeline inside the workflow layer that an investigator actually uses, including FIR drafting, case prioritisation, case linking, and crime analytics. Experiments show that the recognition pipeline more than doubles Rank-1 accuracy over direct sketch matching, while the workflow layer cuts routine paperwork time by a wide margin. Future work will look at on-device inference for field use, identity-preservation losses applied during ControlNet fine-tuning, federated training across police districts so that raw data need not be centralised, and the addition of further biometric modalities such as gait and voice for multi-modal suspect ranking.

REFERENCES

[1] L. Zhang, A. Rao, and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023, pp. 3836-3847.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684-10695.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4690-4699.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815-823.

[5] B. F. Klare, Z. Li, and A. K. Jain, "Matching Forensic Sketches to Mug Shot Photos," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 33, no. 3, pp. 639-646, 2011.

[6] S. Ouyang, T. Hospedales, Y. Song, X. Li, C. Loy, and X. Wang, "A Survey on Heterogeneous Face Recognition: Sketch, Infra-red, 3D and Low-Resolution," Image and Vision Computing, vol. 56, pp. 28-48, 2016.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1125-1134.

[8] M. Hu et al., "Identity-Aware Diffusion for Controllable Face Generation," arXiv preprint arXiv:2408.01233, 2024.

[9] A. Sharma and R. Verma, "Crime Analytics and Predictive Policing using Machine Learning," in Proc. IEEE Int. Conf. Computational Intelligence and Computing Research, 2021.

[10] P. Kumar et al., "A Smart Policing Dashboard for Real-Time Crime Monitoring," in Proc. IEEE Int. Conf. Smart Technologies, 2023.

[11] S. Patel and N. Joshi, "NLP-based Automated FIR Generation for Indian Law Enforcement," in Proc. IEEE Int. Conf. on AI for Governance, 2024.

[12] V. Mirjalili and A. Ross, "Forensic Face Recognition: A Survey," in Handbook of Face Recognition, Springer, 2023, pp. 411-456.

[13] X. Tang and X. Wang, "Face Sketch Recognition," IEEE Trans. Circuits and Systems for Video Technology, vol. 14, no. 1, pp. 50-57, 2004.