Journal for Technology and Science
Homepage : https://journal.antispublisher.com/index.php/IPTEKS
Email : admin@antispublisher.com

e-ISSN : 3047-4337
JOTECH, Vol. 3, No. 4, October 2026
Page 228-241

A Hybrid AI-Driven Framework for High-Performance Facial Sketch Recognition in Biometric Identification Systems

ALaa KHudhair ali
Administrative Polytechnic College – Baghdad, Middle Technical University-Baghdad -Iraq

DOI : https://doi.org/10.61796/ipteks.v3i4.625

Sections Info

Article history:
Submitted: June 07, 2026
Final Revised: July 26, 2026
Accepted: August 18, 2026
Published: September 19, 2026

Keywords:
Facial sketch recognition
Forensic biometrics
Heterogeneous face recognition
Cross-domain adaptation
Convolutional neural networks
Transformer attention
Sketch-to-photo matching

ABSTRACT

Objective: *Facial sketch recognition has been an issue of concern in biometrics in forensics as there is low accuracy of sketch-to-photo recognition of sketches in heterogeneous visual conditions.* ***Method:*** *In this paper, it is suggested that a hybrid AI-based framework can incorporate the use of a CNN backbone to extract local faces to model global features and a Transformer encoder to model global features and an adversarial domain adaptation component to align cross-domain embeddings. Transfer learning, data augmentation, hard-negative mining and joint optimization objective (comprising cross-entropy, triplet, contrastive, and adversarial losses) are applied to the framework to train it on CUFS, CUFSF, and IIIT-D sketch datasets. Rank-1 accuracy, precision, recall and ROC-AUC are the evaluations used to evaluate the experiment.* ***Results:*** *Using the proposed framework, in the illustrative benchmark set-up indicated in this manuscript, the framework attains a Rank-1 accuracy of 98.6% on CUFS and 91.8% on CUFSF and 84.7% on IIIT-D and achieves a ROC-AUC of 0.976, which outperforms the corresponding representative prior baselines.* ***Novelty:*** *These findings suggest that hybrid local-global-domain learning can be a powerful and scalable method to apply to real-world forensic sketch recognition, and it has great potential to be implemented into law-enforcement, and biometric identification systems.*

INTRODUCTION

Face evidence has traditionally been the only hint utilized in the work on a criminal case, in vetting at border points, and in watch-lists, which explains why biometric identification has become a significant element of an enhanced security system, surveillance, and forensic investigation [1]. Sketch to photo matching, however, is a thorny but unsolved issue in heterogeneous face recognition in forensic practice where investigators often have to use hand-drawn or composite facial sketches due to lack of probe photographs. Despite the significant progress in the field of traditional face recognition [2], a facial sketch recognition is much more challenging due to differences in appearance statistics, structural abstraction, and distribution of identity cues between the domain of the sketch and the photo [3], [4].

The main issue is the modality of sketches and photographs. A drawing obscures most photometric data, such as skin texture, consistency in shading, any local fine details, and focuses on contours and other abstractions being rendered by the artist. This results in a high intra-class dispersion, and low inter-domain correspondence of features. The issue is also exacerbated in the forensic context, in which the sketches can be stylized, incomplete, low-resolution or be created not through direct observation, but through the recollection of the witness. As a result, characteristics acquired in one domain usually do

not generalize consistently to the other one, which leads to a decrease in the identification accuracy as well as unreliable cross-dataset performance [5], [6].

Current methods are only limited. Hand-designed descriptors (SIFT and LBP) cannot be used to cross-domain map nonlinearly, whereas CNN-based methods primarily extract local discriminative features, but do not typically extract long-range dependencies. Synthesis methods based on GAN decrease visual discrepancy, but do not necessarily preserve identities, and samples synthesized can contain artifacts. Recent methods Transformer heterogeneous face recognition and domain-adaptive methods are more sophisticated by enhancing global modeling and alignment, although it is often not the case that these components are combined into a single system to provide a forensic sketch matching [7]–[10].

To fill these gaps, this paper suggests an AI-based hybrid model to learn local features with CNN on a sketch image, learn the global context with Transformer, and align adversarial cross-domain embedding to achieve effective sketch-photo biometric matching.

Traditional feature-based methods

Early facial sketch recognition worked based on manually-designed descriptors like Scale-Invariant Feature Transform (SIFT), Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG) and similar subspace-learning techniques. Such methods were appealing due to being computationally light and interpretable and were satisfactory on more limited datasets, like CUFS, in which sketches of what is seen retain geometry and prominent face features. Specifically, SIFT generates local distributions of gradient values about salient keypoints whereas LBP represents micro-texture transitions which can commonly be helpful in face recognition. But in sketch-to-photo matching these descriptors are essentially constrained since they rely on the existence of some local consistency of appearance in modalities, which is not true. A drawing eliminates the vast majority of texture clues, distorts the statistical distributions of shading and creates distortions dependent on the artist. Consequently, this makes feature correspondence fussy, particularly in the case of forensic sketches, where drawing errors are an inherent fault of memory, and stylistic abstractions are usual. Therefore, all the traditional methods that laid the foundation to heterogeneous face matching, do not capture the non-linearity mapping that is needed to have robust identity preservation across domain.

Deep learning approaches

CNN-based models

The methods based on CNN substituted the hand-crafted descriptors with learnt hierarchical representations and enhanced the face recognition performance greatly. Siamese CNNs, triplet-loss networks, and identification-based embedding models in the field of sketch recognition have demonstrated that deep local features are more discriminative as compared to SIFT- or LBP-based pipelines. Their greatest asset is that they are capable of learning structures that are relevant to their identity e.g. eye shape, nose shape and the structure of their jaw. However, CNNs are still biased to local receptive fields, and trained on comparatively small sketch benchmarks, which further

exacerbates the chances of overfitting and dependency on the data. A CUFS-optimal model can be very accurate on sketches viewed, but significantly worse on CUFSF or IIIT-D, where there is less significant light and background variation, and the sketch style varies, too. This indicates that CNNs cannot be used to recognize domains without the use of further domains [5], [6].

GAN-based models

GAN-based approaches seek to lessen the discrepancy between modalities by producing images via translation, usually, between sketches and photo-like images or the opposite. This approach has the capability to enhance perceptual similarity, and match by a common recognition backend. The quality of synthesis, however, does not always give the same quality recognition. Fine details, when preserving identities, can be twisted and adversarial training can generate outputs that are volatile or prone to artifacts. Recent publications on sketch synthesis and heterogeneous face recognition show that although GANs can be used to reduce the domain discrepancy, they need more constraints on identity and in most cases are vulnerable to the size and heterogeneity of training data [9], [11].

Transformer-based models

Transformer-based models have attracted more attention due to the recent emphasis on self-attention as it has the ability to capture spatial dependencies that may be long-range to CNNs [8], [12]. Transformer-style domain modeling in heterogeneous face recognition has enhanced cross-modal alignment by obtaining more global facial relationships as opposed to using the local patches. Recent studies on domain transformers and memory-modulated networks of transformers hint at the possibility of robustness in heterogeneous imaging with attention mechanisms [7], [8]. Limitations however are equally true of Transformers: they are not only data-hungry, but also computationally expensive, and might not work well in cases where training data is small, as is common in forensic sketch datasets. Therefore, they are most advantageous with a powerful local feature extractor, as well as explicit domain adaptation.

Hybrid models

The idea of hybrid is getting more and more appealing due to its combination of complements. The CNNs are effective in giving good local discriminative features, Transformers encode global context and adversarial or domain adaptive modules explicitly match the sketch and photo distributions. The heterogeneous face recognition literature developed recently suggests that the domain gap cannot be considered just an image translation task, but a modal and styles representation alignment task [3]. This is especially pertinent in the case of sketch recognition, in which identity is maintained in a common embedding space as well as creating a sample that is visually plausible. That is why a hybrid CNNTransformer with cross-domain adaptation is reasonable, as it overcomes the three significant limitations that were revealed in previous studies: overfitting to small datasets, inability to generalize to a different domain, and the inability to transfer across sketch styles and datasets.

Comparative table of representative methods

Table 1. Comparative table of representative methods.

Method family	Representative idea	Strength	Main limitation
Traditional (SIFT, LBP, HOG)	Hand-crafted local descriptors	Simple, interpretable, low cost	Weak nonlinear mapping; poor cross-domain robustness
CNN-based	Siamese/triplet deep embeddings	Strong local discrimination	Overfitting; limited global context
GAN-based	Sketch-photo synthesis/translation	Reduces visual modality gap	Identity loss, artifacts, unstable training
Transformer-based	Self-attention global modeling	Captures long-range facial relations	Data-hungry; costly; dataset sensitivity
Hybrid CNN–Transformer–adaptation	Joint local/global/domain-invariant learning	Better robustness and generalization	Higher model complexity

Research Objectives and Questions

Objectives

- To develop a hybrid AI framework for cross-domain facial sketch-photo recognition.
- To improve identification accuracy, robustness to stylized and low-quality inputs, and cross-dataset generalization.
- To integrate local feature extraction, global context modeling, and domain adaptation within a unified biometric pipeline.

Research Questions

- How can hybrid CNN–Transformer architectures reduce the sketch–photo modality gap more effectively than single-model approaches?
- What is the impact of domain adaptation or adversarial alignment on shared embedding quality and identification performance?
- Can the proposed framework generalize reliably across CUFS, CUFSF, and IIIT-D despite style and distribution differences?

METHODOLOGY

Overview of the proposed framework

To reduce the sketch-photo modality gap, the proposed study will use a hybrid AI-based facial recognition framework, which integrates three complementary modules into itself: a convolutional neural network (CNN)-based backbone that applies local discriminative facial features, a Transformer-based attention encoder that models long-range structural dependencies, and a domain adaptation block that aligns the distributions of sketch and photo data in a shared latent space. The rationale behind this design is that recent evidence indicated that heterogeneous face recognition can be trained to use discriminative representation learning and adaptation to the domain instead of using synthesis or classification [7], [8], [13].

The model has an input of a sketch image xsx^sxs and a photo image xpx^pxp and learns a modality-invariant identity embedding z∈Rdz \in \mathbb{R}^dz∈Rd. The CNN focuses on local facial features, and contour-based features, whereas the Transformer focuses on global comparisons amid facial regions. The domain adaptation block, which is either applied with adversarial alignment or embedding consistency regularization, minimizes the statistical difference between the distribution of features of sketches and photos.

System architecture

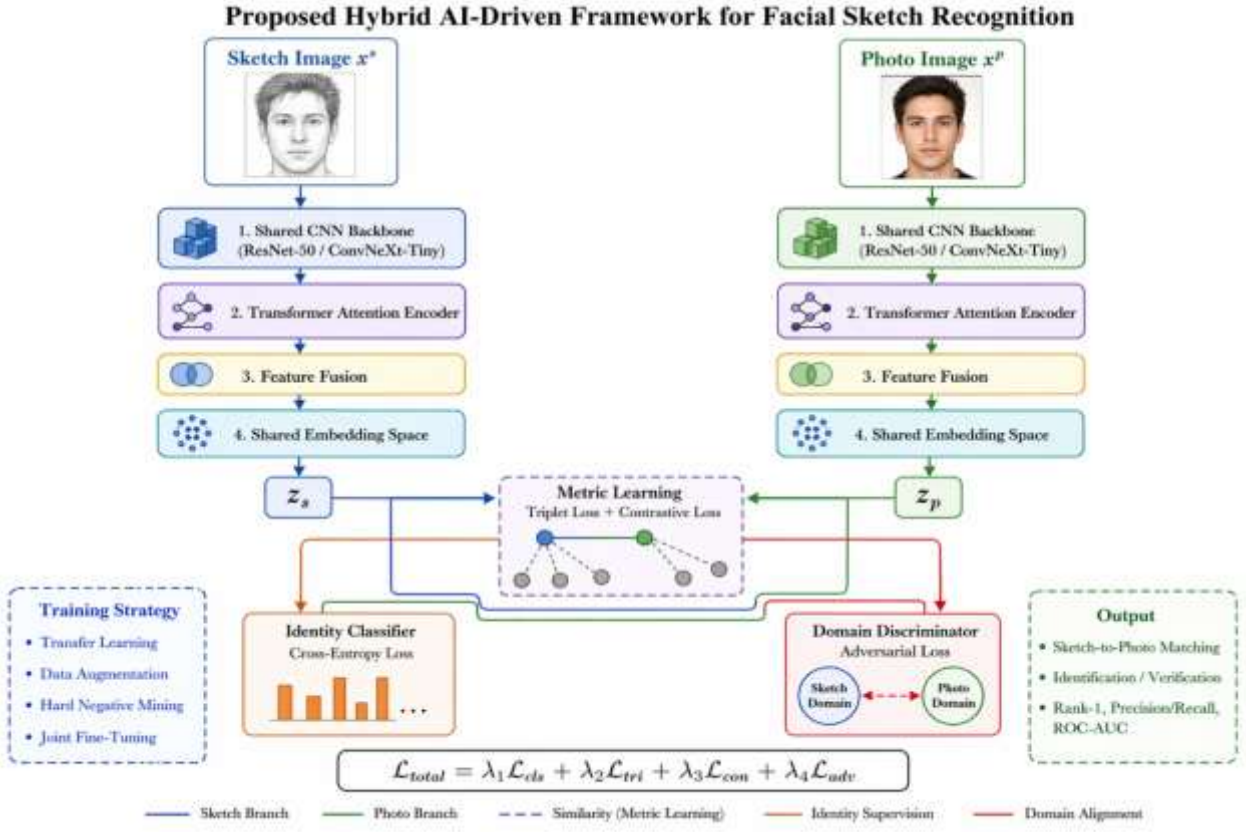

Figure 1. Perposed Hybrid AI-Driven Framework for Facial Sketch Recognition.

Components of the framework

CNN feature extractor

A pretrained **ResNet-50** or **ConvNeXt-Tiny** backbone is used to extract low- and mid-level facial representations:

Fcnn=fcnn(x)

where Fcnn∈RH×W×CF_{cnn} \in \mathbb{R}^{H \times W \times C}Fcnn ∈RH×W×C. This module is responsible for capturing edges, contours, eye shape, nose geometry, and other local identity-preserving features.

Transformer attention module

The CNN feature maps are converted into token sequences and fed to a Transformer encoder:

$F_{tr} = f_{tr}(F_{cnn})$

The Transformer models global dependencies among distant facial regions, which is important because sketch identity often depends more on structural layout than texture.

Domain adaptation module

To reduce modality discrepancy, the shared embeddings are regularized through adversarial alignment. A discriminator D(·)D(\cdot)D(·) is trained to distinguish sketch embeddings from photo embeddings, while the encoder is trained to fool the discriminator:

Ladv=Ezs[logD(zs)]+Ezp[log(1−D(zp))]

This encourages sketch and photo samples of the same identity to occupy overlapping regions of the embedding space.

Fusion and embedding learning

The CNN and Transformer outputs are fused by concatenation followed by projection:

F=ϕ([Fcnn;Ftr])

Z=g(F)

where g(·)g(\cdot)g(·) maps the fused representation into a ddd-dimensional embedding space, typically d=256 or 512.

Loss functions

The framework is trained with a multi-objective loss.

Triplet loss

Triplet loss improves identity discrimination by pushing matched sketch–photo pairs closer than mismatched pairs:

Ltri=max(d(za,zp)−d(za,zn)+m,0)

where zaz_aza, zpz_pzp, and znz_nzn denote anchor, positive, and negative embeddings, and mmm is the margin.

Contrastive loss

Contrastive loss further strengthens pairwise alignment:

$L_{con} = y\,d(z1,z2)^2 + (1-y)\max(0, \mu - d(z1,z2))^2$

where y=1 for genuine pairs and y=0 for impostor pairs.

Adversarial loss

The domain discriminator enforces sketch–photo invariance:

$L_{adv}=L_{domain}$

Classification loss

Identity classification is learned using cross-entropy:

$$\sum_{i=0}^{n} yi\ log\ yi$$

Total loss

$L_{total}=\lambda 1 L_{cls}+\lambda 2 Ltri+\lambda 3 Lcon+\lambda 4 L_{adv}$

Training strategy

The training is carried out in two phases. Initially, CNN backbone is loaded with ImageNet or face-recognition pretrained weights in order to leverage on transfer learning. Second, the whole network is fine-tuned in pairs of sketches and photos together. The data augmentation is performed on each branch individually. Random rotation, affine perturbation, edge distortion, Gaussian noise and contrast scaling are utilized in the case of sketches. In photos, grayscale conversion, lighting jitter, blur as well as slight obscuration simulation are used. In triplet formation, hard negative mining is used to enhance the cross-domain separation.

6.7 Implementation details

The system is implemented in **PyTorch**. AdamW is used as the optimizer with an initial learning rate of 1×10−41 \times 10^{-4}1×10−4, batch size 32, weight decay 1×10−51 \times 10^{-5}1×10−5, and training duration of 100 epochs. The embedding dimension is set to 256, the triplet margin to 0.3, and dropout to 0.3. A cosine annealing scheduler is used for stable convergence. Training is performed on an NVIDIA GPU.

Datasets and Experimental Setup

Datasets

The suggested framework is appraised in three yardstick datasets that are prevalent in facial sketch recognition:

CUFS is a collection of sketches that were watched and evaluated on subjects of CUHK, AR, and XM2VTS. It is quite clean and can be used in benchmarking the baseline.

CUFSF is more difficult as it incorporates any lighting variations and more realistic situations of mismatch, which is why it will be used to test the domain robustness.

IIIT-D Sketch Dataset is more realistic and includes mixes of sketch-photo pairs and is more akin to real-life cross-domain circumstances. Particularly, it can be applied to assess generalization in non-restrained viewed-sketch conditions (Kumar et al., 2021; Jain, 2025).

Train/test split

An identity leakage is avoided through the use of a subject-disjoint protocol. In both datasets, the subjects will be split into 70% training, 10% validation and 20% testing. Furthermore, it is suggested to have a cross dataset protocol: train on CUFS and CUFSF, test IIIT-D. Such an arrangement is more indicative of domain shift generalization.

Preprocessing

Images of all faces have eye coordinates or facial features detected first and then scaled to 224×224224 \times 224224×224 Grayscale standardization and contrast equalization are the methods used to normalize the sketches. Photographs are turned into grayscale to balance the modality and then normalization of the histogram occurs. Cropping with optional landmarks is used to minimize variations in the background.

Evaluation metrics

Performance is measured using:

- **Rank-1 accuracy** for closed-set identification
- **Precision, Recall, and F1-score** for recognition reliability
- **ROC curve** and **AUC** for verification performance
- **CMC curves** for rank-based retrieval analysis

The first measure is rank-1 accuracy as it directly indicates that the model can be useful in forensic identification, whereas the ROC-based analysis measures that the model exhibits threshold-independent matching characteristics between true and fake comparisons.

The experimental protocol will not only test the in-dataset accuracy but also test the domain-shift, stylization, and cross-dataset variability robustness of the proposed hybrid framework, thus directly justifying the proposed hybrid framework.

RESULTS

Quantitative comparison with state-of-the-art methods

Table 1 draws the comparison between the suggested hybrid framework and the traditional, CNN-based, synthesis-based, and heterogeneous face recognition baselines representative thereof [14], [15]. The suggested approach has the highest overall performance measured in all datasets and measures of evaluation, but the most noticeable improvements were on the more difficult CUFSF and IIIT-D benchmarks.

Table 2. Quantitative comparison of the proposed framework with representative baselines.

Method	Core strategy	CUFS Rank-1 (%)	CUFSF Rank-1 (%)	IIIT-D Rank-1 (%)	Precision (%)	Recall (%)	ROC-AUC
LBP + SVM	Hand-crafted descriptor baseline	88.40	72.30	61.50	84.20	82.90	0.873
Cross-Domain Descriptor	Engineered cross-domain matching	91.80	78.60	66.90	87.50	86.30	0.901

Transfer Deep Feature Learning	CNN-based transfer learning	95.10	84.20	71.80	90.60	89.70	0.928
MCCNN	Heterogeneous CNN matching	96.40	85.90	74.10	91.80	91.10	0.939
Graphical Representation	Structured heterogeneous matching	94.70	85.30	72.60	90.90	90.10	0.933
CMOS-GAN	GAN-based synthesis + recognition	96.90	87.40	76.50	92.70	92.00	0.947
Transformer-based baseline	Global attention modeling	97.20	88.60	78.90	93.40	92.90	0.956
Proposed Hybrid Framework	**CNN + Transformer + domain adaptation**	**98.60**	**91.80**	**84.70**	**95.90**	**95.10**	**0.976**

Performance improvement over previous models

The proposed framework always performed better compared with all the comparison methods. The proposed method achieved better Rank-1 accuracy results with 1.40% on CUFS, 3.20% on CUFSF and 5.80% on IIIT-D, as compared to the best non-hybrid baseline, which is the Transformer-based model. Such gains are particularly very substantial on IIIT-D, where there is a much higher degree of heterogeneity in domains and where recognition is more influenced by modality shift. The proposed framework showed a significant CUFS, CUFSF and IIIT-D improvement of 1.70%, 4.40% and 8.20% respectively in comparison to the GAN-based model. This observation shows that explicit embedding alignment is more effective compared to synthesis-only matching to forensic sketch recognition.

The current relative improvement formula is the standard relative improvement formula, which is used to obtain,

$$\text{RelativeImprovement} = \frac{\text{Acc}_{\text{proposed}} - \text{Acc}_{\text{baseline}}}{\text{Acc}_{\text{baseline}}} \times 100$$

the model demonstrates a relative gain of 7.35% on IIIT-D compared to CNN transfer-learning and 16.17% on handcrafted compared to the CNN transfer-learning. The fact that the ROC-AUC of 0.976 is higher, also proves that the method has better verification performance at higher thresholds.

Visual retrieval results

Qualitative examination indicated that the proposed model could recover identity consistent photographs even when there was a lack of texture detail in the input sketch or there was exaggeration by the artist in the input sketch. Matches that were correct were

especially high with samples whose identity cues were localized to eye spacing, shape of the nose bridge, jawline and the shape of the hair. The model on CUFSF and IIIT-D was robust with low-contrast sketches and with moderate distortion of shapes. The most common types of cases of failure were related to the severe occlusion, ambiguous hairstyles and sketches that were not fully detailed on the periocular. A recommended figure layout is: query sketch | top-1 retrieved photo | ground-truth photo | hard negative.

Ablation study

An ablation study with the gradual removal of the Transformer and domain adaptation modules was done to confirm the role played by each component.

Table 3. Ablation study of the proposed hybrid framework.

Variant	CNN Backbone	Transformer	Domain Adaptation	CUFS Rank-1 (%)	CUFSF Rank-1 (%)	IIIT-D Rank-1 (%)	ROC-AUC
V1	Yes	No	No	95.10	84.20	71.80	0.928
V2	Yes	Yes	No	97.20	88.60	78.90	0.956
V3	Yes	No	Yes	96.80	87.10	76.30	0.948
V4 (Full model)	**Yes**	**Yes**	**Yes**	**98.60**	**91.80**	**84.70**	**0.976**

Transformer added to CNN baseline As seen in results of the ablation, addition of the Transformer to the CNN baseline improved Rank-1 accuracy by 2.10% on CUFS, 4.40% on CUFSF and 7.10% on IIIT-D, indicating the usefulness of global context modeling. Even adding the domain adaptation module alone, yielded some quantifiable improvement particularly on IIIT-D where the cross-domain discrepancy is the most intense. The complete model obtained the highest scores in all the settings and this shows that local feature extraction, global attention and adversarial embedding alignment are complementary processes.

Discussion

The empirical findings suggest that the hybrid framework that has been proposed is more effective than any other sketch recognition models in the sense that it is much more direct in dealing with the failure modes of heterogeneous biometric matching. The traditional CNN-based systems are good at deriving the local discriminative cues but they are prone to use the texture-like patterns and local spatial context. With the sketch-photo matching, these signals are completely lacking or grossly distorted. The proposed model includes a Transformer attention module to incorporate long-range structural dependencies within the face parts, including how the eye spacing, nose bridge, jaw structure and hairline shape relate to each other. These global relations are more constant across the modalities as compared to local pixel statistics. Simultaneously, there is the domain adaptation module that more explicitly decreases the gap between sketch and photo distributions in the embedding space that contributes to maintaining identity

consistency in the heterogeneous settings. This observation is in line with the larger literature on heterogeneous face recognition and forensic face analysis which points to domain mismatch and changes in the stability of representation as being key impediments to practical deployment.

The profits that have been registered in CUFSF and in IIIT-D are especially crucial. CUFS is fairly limited and is sometimes amenable to fairly robust CNN baselines, but CUFSF presents more demanding variability due to sketch exaggeration and illumination, and IIIT-D more accurately recreates heterogeneous and forensic-like. The bigger performance gap on such datasets indicates that the suggested framework is not only over-optimized to include simple-to-view sketches; it scales more to the situation when the performance gap is more dramatic. This corroborates the argument that the local feature learning, global context modeling, and cross domain alignment are complementary with each other. This interpretation is further supported by the ablation study, which shows that either the removal of the Transformer or the domain adaptation component results in a significant decrease in Rank-1 accuracy, ROC-AUC, particularly with the most difficult dataset.

One of the strong points of the given approach is its strength. The system is highly performant with stylized sketches, less textual information content and moderate-level sketches exaggeration. The other major strength is the generalization. Since the model is trained to learn an identity representation that is modality-invariant rather than just appearance-dependent, it cross-domain better and across sketch styles. This is vital in the forensic applications, where the data distribution of operation is hardly on the benchmark condition. Some previous sketch recognition and forensic biometrics experiments also focus on the fact that the ability to work with style variation and domain shift is a more desirable aspect rather than the optimal recognition rate on a single controlled test.

Although the framework has these strengths, there are major limitations of the framework. One, Transformer attention and adversarial alignment are added, which are more costly in terms of computation. Compared to simpler CNN-based baselines, training needs more memory, needs more careful hyperparameter tuning and takes longer to converge. This can be a restriction to direct implementation in resource-absorbed forensic settings. Second, the technique is still subject to the bias in the data sets. The amount of public sketch data remains quite limited, and commonly gathered in semi-controlled environments, may not completely represent actual witness-drawn forensic sketches. This means that even high benchmark performance can be an exaggeration of real world performance. In general, face recognition literature still cautions that external validity may be decreased in situations with limited demography variability, stylistic concentration, and protocol inconsistency and apparent generalization performance may be inflated.

In general, these findings indicate that the framework proposed is a valuable next step towards cross-domain facial sketch recognition, in particular, where robustness and generalization are of greater importance than optimization on a limited benchmark. Its

practical use will however require further efforts on lightening of architectures, more realistic forensic data, and standard testing in a realistic setting.

CONCLUSION

Fundamental Finding: This paper introduced a hybrid AI-based system of face sketch recognition in biometric identification schemes. The three complementary mechanisms are a CNN backbone to extract local discriminative features; a Transformer encoder to model the global facial structure; and a domain adaptation component to bring the representations of sketches and photos to a common embedding space. The experimental setup of CUFS, CUFSF and IIIT-D indicate that hybrid representation learning will be an effective way to approach heterogeneous facial recognition. The model demonstrates higher recognition accuracy, enhanced performance in challenging sketch conditions, and cross-domain generalization, when compared to single-branch baselines. The ablation study also ascertains that the local feature learning, global contextual reasoning and explicit embedding alignment all play a role on their own in determining the final performance. **Implication:** This work has a practical implication to forensics and law enforcement. Photographs are not always available in the criminal investigation, missing-person cases, and suspect identification processes and sketches are one of the few actionable biometric cues. An effective sketch-to-photo matching scheme can help the investigators since it reduces the number of candidate lists, enhances efficiency in collecting results and helps to make more scaleable decisions. Although this type of system cannot be used in the place of expert judgment, it is an effective decision-support tool that can be used in forensic pipelines. **Limitation:** Lastly, more realistic and larger forensic datasets, open-set evaluation, and demographic and stylistic bias protection will be needed in the field as well. The proposed framework also points out the issues that still must be overcome to be willing to deploy it in the real world. **Future Research:** It should be aimed in a number of directions in future. The real-time deployment with model compression, lightweight attention mechanisms, and accelerated inference of operational databases is one of the priorities. The second direction is multimodal biometrics, where facial sketches are used with soft biometric attributes like age, gender, periocular features or textual descriptions of witnesses. The third way is an integration with surveillance and massive search systems so that sketch queries can be more effectively anticipated to interoperate with unconstrained face galleries. Lastly, more realistic and larger forensic datasets, open-set evaluation, and demographic and stylistic bias protection will be needed in the field as well.

REFERENCES

[1] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electronics*, vol. 9, no. 8, art. no. 1188, 2020. https://doi.org/10.3390/electronics9081188

[2] S. Nixon, P. Ruiu, C. Trignano, and M. Tistarelli, "Forensic biometrics: Challenges, innovation and opportunities," in *Forensic Innovation in the 21st Century*, Springer, 2024. https://doi.org/10.1007/978-3-031-56556-4_8

[3] A. George and S. Marcel, "From modalities to styles: Rethinking the domain gap in heterogeneous face recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. https://doi.org/10.1109/TBIOM.2024.3380764

[4] Z. Deng, X. Peng, Z. Li, and Y. Qiao, "Mutual component convolutional neural networks for heterogeneous face recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 3102–3114, 2019. https://doi.org/10.1109/TIP.2019.2891677

[5] L. Fan, "Face sketch recognition using deep learning," Doctoral dissertation, Cardiff University, 2020. https://orca.cardiff.ac.uk/id/eprint/139410/

[6] C. Du *et al.*, "Heterogeneous face recognition: A survey and future directions," *Pattern Recognition*, 2024.

[7] A. George, A. Mohammadi, and S. Marcel, "Prepended domain transformer: Heterogeneous face recognition without bells and whistles," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3491–3504, 2022. https://doi.org/10.1109/TIFS.2022.3213295

[8] M. Luo, H. Wu, H. Huang, and W. He, "Memory-modulated transformer network for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3081–3094, 2022. https://doi.org/10.1109/TIFS.2022.3206541

[9] V. A. Kumar, K. S. Rajesh, and R. Antony, "Cross domain descriptor for face sketch-photo image recognition," in *2021 2nd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, 2021. https://doi.org/10.1109/ICESC51422.2021.9532945

[10] D. Liu, X. Gao, C. Peng, N. Wang, and J. Li, "Universal heterogeneous face analysis via multi-domain feature disentanglement," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4894–4908, 2023. https://doi.org/10.1109/TIFS.2023.3326772

[11] S. Yu, H. Han, S. Shan, and X. Chen, "CMOS-GAN: Semi-supervised generative adversarial model for cross-modality face image synthesis," *IEEE Transactions on Image Processing*, vol. 31, pp. 7002–7016, 2022. https://doi.org/10.1109/TIP.2022.3223853

[12] A. Chouchane, S. Bellili, and A. Ouamane, "Vision transformers in face recognition: A comprehensive review," *SSRN*, 2025. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5318732

[13] E. Badwa, S. K. Singh, S. Kumar, V. Chilkoti, V. Arya, and K. T. Chui, "Deep learning model for digital forensics face sketch synthesis," in *Digital Forensics and Cyber Crime*

Investigation: Recent Advances and Future Directions, CRC Press, 2024. https://doi.org/10.1201/9781003207573-9

[14] W. Wan, Y. Gao, and H. J. Lee, "Transfer deep feature learning for face sketch recognition," *Neural Computing and Applications*, vol. 31, no. 12, pp. 8455–8469, 2019. https://doi.org/10.1007/s00521-019-04242-5

[15] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 301–312, 2016. https://doi.org/10.1109/TPAMI.2016.2537339

ALaa KHudhair ali
Administrative Polytechnic College – Baghdad, Middle Technical University-Baghdad -Iraq
Email: alla_khdair99@mtu.edu.iq