SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I8P125

Volume 11 Issue 8, 260-268, August 2024
© 2024 Seventh Sense Research Group®

Original Article

Heterogeneous Sketch-Face Photo Recognition in Forensic Science Laboratories

Devendra A. Itole[1], M. P.Sardey[2], Milind P. Gajare[3]

[1,2,3]*Department of Electronics and Telecommunication AISSMS IOIT, Pune, Maharashtra, India.*

[1]*Corresponding Author : devendra.itole@aissmsioit.org*

Received: 19 June 2024 Revised: 31 July 2024 Accepted: 15 August 2024 Published: 31 August 2024

Abstract - *This research presents a pioneering framework, termed X-Bridge, aimed at automating the identification of diverse faces through facial sketches. The significance of this framework lies in its potential applications in security and surveillance domains. The study advances the field by a. Conducting an in-depth analysis of conventional neural network architectures utilized for image classification, particularly focusing on their effectiveness in facial recognition tasks. b. Investigating the latest parameters essential for accurate facial recognition and their integration into various neural network structures to enhance performance. c. Assessing potential cross-modal connections that could facilitate more robust facial recognition systems.d. Introducing a novel Generative Adversarial Network (GAN)-based strategy, X-Bridge, specifically tailored to surpass existing standards on a meticulously curated dataset dedicated to facial recognition. Through these endeavors, the X-Bridge framework exceeds current benchmarks, demonstrating its efficacy in automating facial recognition tasks. This research contributes to the advancement of automated facial recognition technology, offering promising implications for security and surveillance applications.*

Keywords - *Cross-modal bridge, Heterogeneous face identification, Image-to-sketch conversion, Machine learning, Artificial neural networks, Categorization, Validation, Identifying.*

1. Introduction

Face Recognition (FR) denotes the process of person verification or identification based on facial features extracted from an image or video source. FR stands out as a prominently explored domain within computer vision, garnering substantial attention over recent decades owing to its multifaceted applications. Foremost among these applications is its pivotal role in biometrics[1]. Unlike other biometric modalities such as fingerprints or iris scans, FR possesses the unique capability to identify subjects non-intrusively without requiring their active participation[2]. This renders it invaluable for diverse purposes, including security systems, forensic investigations, and the identification of individuals within crowded environments. Furthermore, FR serves as an additional layer of security in authentication systems. Beyond biometrics, FR finds utility in domains like gender classification, emotion recognition, database searching, and witness identification, among others[3][4].

Notwithstanding its widespread adoption, FR remains a formidable technical challenge due to a myriad of external and internal factors. External conditions like variations in illumination, pose, or occlusion, coupled with internal factors such as facial expressions and aging, contribute to the complexity of FR algorithms[5].

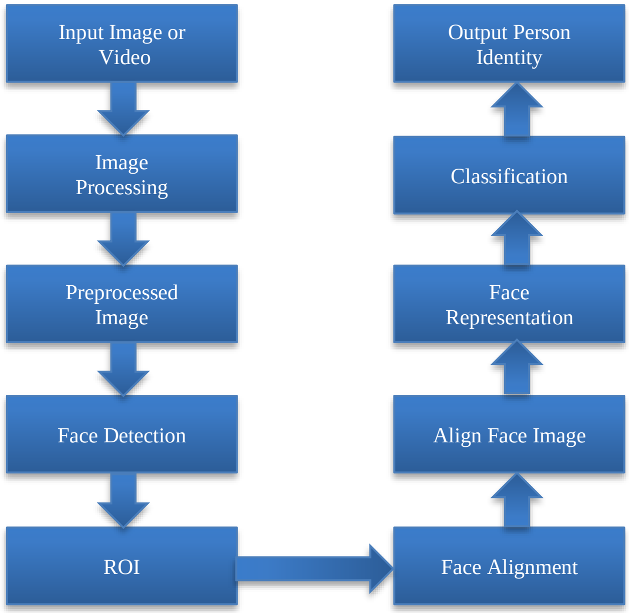

Fig. 1 Face recognition process

1.1. Image Preprocessing

It involves eliminating unwanted distortions (noise) from an image while preserving vital information. Among the techniques are geometric conversions, brightness and color adjustments, local manipulations (such as gradients operator and filters), and analysis of frequencies.[6][7]. Although preprocessing reduces information, it has often been included in Face Recognition (FR) systems to enhance performance, leveraging prior knowledge about the image[8].

1.2. Face Detection

This process identifies human faces within an image. Various algorithms exist, with Haar cascade detection being one of the most renowned.

1.3. Face Alignment

Following face detection, this step further processes the Region of Interest (ROI). It may employ advanced preprocessing techniques based on prior knowledge about the expected presence of a human face, aiming to mitigate challenges like pose or non-rigid expression. While not mandatory, this step enhances FR accuracy[9][11].

1.4. Face Representation

This entails extracting features or taking an accurate picture of the facial image. Elastic Bunch Graph Matching, Principal Component Analysis, Fisher Linear Discriminant Analysis, Neural networks, and 2D and 3D facial synthesis are outstanding notable methods in the field of facial recognition and synthesis.[1][12].

1.5. Classification

This process assigns a new observation to a specific category. In FR, it determines a person's identity or verifies whether they are who they claim to be. For this, techniques for classification like Neural Networks (NEUs), Supported Vector Machines (SVM), and Bayesian classification algorithms are frequently used.[13].

Heterogeneous Face Recognition (FR) is a specialized form of FR that operates across distinct visual domains. Such approaches find crucial applications, particularly in security and surveillance[14][15]. For instance, in law enforcement, heterogeneous FR aids in identifying individuals based on sketches derived from eyewitness descriptions. Another application lies in FR using infrared light, offering the advantage of visibility in low-light conditions[16]. This technology is instrumental in security systems where standard RGB cameras are ineffective due to inadequate lighting[17].

FR can be viewed as a subset of object recognition, a challenging task due to its nonlinear nature. The difficulty arises from the inherent similarity among human faces and their non-rigid nature. Variations in facial appearance stem from both internal factors and external conditions.

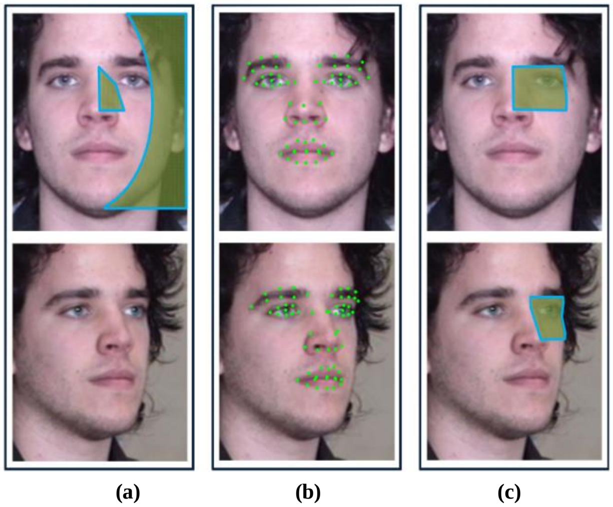

Fig. 2 Pose fluctuation presents problems for FR (a) loss of semantic correlation, (b) self-occlusion, and (c) nonlinear warping of face textures

Internal sources of facial variation are inherent physical attributes that remain unaffected by external observers. These can be classified into intrapersonal and interpersonal attributes.

Intrapersonal attributes pertain to differences within an individual's appearance, such as facial expressions, aging, hairstyle variations, and accessories like glasses. Interpersonal attributes, on the other hand, account for differences in appearance between different individuals, including gender, ethnicity, and age[18][19].

External factors can influence how a face looks by interacting with lights or the relative positioning of the facial features and the observer[1]. Pose, scale, occlusion, lighting fluctuations (illumination), and imaging characteristics (resolution, noise, focus, and image domain) are all included in these circumstances. Pose changes in FR provide three distinct challenges: nonlinear warping of facial textures, loss of semantic congruence, and self-occlusion[20].

While interpersonal attributes are desirable for FR, intrapersonal differences and external conditions present significant challenges. In many cases, variations induced by intrapersonal differences and external conditions can overshadow interpersonal differences within standard subspaces, exacerbating the complexity of FR tasks[21].

It is noteworthy that the utilization of surveillance systems raises concerns regarding citizen privacy. The deployment of systems incorporating face detection and recognition capabilities may potentially infringe upon individuals' privacy rights by enabling monitoring of their movements and actions[24].

2. Literature Survey

Gulrajani et al. (2017) focused on improving the training of Wasserstein GANs (WGANs) for better stability. Larsen et al. (2016) explored using learned similarity metrics in autoencoders beyond pixels, contributing to better reconstruction in image generation. Gruber (2018) examined the use of VAEGANs for generating facial images. Mirza and Osindero (2014) introduced conditional GANs, allowing for controlled image generation based on input conditions. Kingma and Welling (2013) proposed the VAE framework for learning latent variable m.

2.1. Problem Definition

The primary problem is to improve the accuracy and realism of facial image generation and recognition, focusing on GANs, VAEs, and autoencoders.

2.2. Motivation and Application

Motivated by the need for robust face recognition systems in security, surveillance, and user authentication applications.

2.3. Face Recognition Datasets

Gross et al. (2010) introduced the Multi-PIE dataset, which is widely used for face recognition research. Gao et al. provided a large-scale Chinese face database for benchmarking facial recognition models. Huang et al. benchmarked video-based face recognition systems using the Cox face database.

2.4. Network Architectures

Schroff et al. (2015) presented FaceNet, a unified embedding model for face recognition and clustering. Deng et al. (2018) proposed ArcFace, which uses additive angular margin loss for enhanced face recognition accuracy.

2.5. Loss Function

Wang et al. (2018) introduced a large-margin cosine loss function for deep face recognition. Wang et al. (2017) focused on L2 hypersphere embedding with NormFace for face verification. Wang et al. (2018) proposed an additive margin softmax loss for better discriminative face verification.

2.6. GAN

Mirza Osindero (2014) provided the foundation for conditional GANs, allowing for the generation of images with specified attributes. Gulrajani et al. (2017) improved WGAN training for more stable and realistic image generation.

2.7. Cross-Modal Bridge

2.7.1. Dimensionality Reduction

Hadsell et al. (2006) focused on learning invariant mappings for dimensionality reduction, a technique crucial for bridging different modalities.

2.8. Feature Extractor

2.8.1. DeepFace

Taigman et al. (2014) introduced DeepFace, aiming to close the performance gap between humans and machines in face verification.

2.8.2. Joint Identification-Verification

Sun et al. (2014) developed a deep learning framework combining identification and verification for robust face recognition.

2.9. Pipeline of the System

Systems typically involve face detection, feature extraction, and classification stages, with GANs or VAEs used for image generation or enhancement.

2.10. Facial Features Preservation Score

2.10.1. VAEGAN

Gruber (2018) used VAEs to generate facial images while preserving key facial features, leading to higher preservation scores in generated images.FaceNet & ArcFace: Emphasis on embedding techniques that maintain the integrity of facial features across transformations.

3. Methodology

Training Set: A group of information utilized to train a model for machine learning, consisting of labeled instances from which the algorithm trains.[4][10].

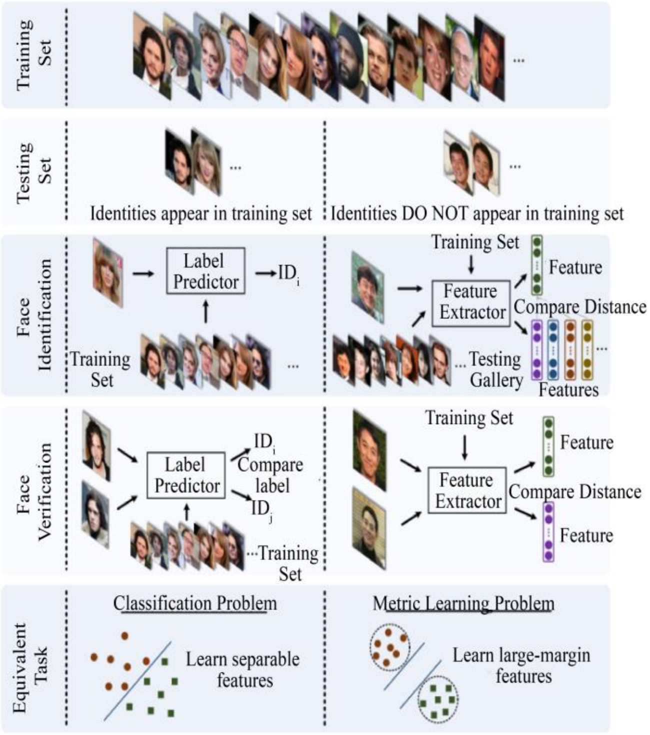

Fig. 3 Comparing open-set and closed-set recognition of faces methods [1]

3.1. Testing Set

A different collection of information is utilized to evaluate the effectiveness of a trained machine learning model, enabling it to evaluate its capacity to generalize unknown information.[2][27][29].

3.2. Face Identification

The method of establishing an individual's identification from a given photograph by comparing it to a database containing existing pictures[16].

3.3. Face Verification

Confirming whether two face photos are of the same people or not is usually done for control of access or validation.[6][30][31].

Face Recognition Modern machine learning relies heavily on datasets to train and validate categorization techniques[8][14]. It offers an examination of well-known datasets for sketch-based and face recognition. A comparison between these datasets is presented in Table 1. Notably, there has been a significant performance gap among methods due to the use of private datasets from companies like Google, Facebook, and Microsoft. However, the availability of newly accessible datasets with millions of images helps bridge this gap. Openly accessible datasets, as well as obstacles, significantly enhance the reproducibility of work[1][32].

The table provides a summary of various face recognition datasets, listing the number of pictures/videos and Identities (IDs), conditions (e.g., laboratory or variable), and resolution[6][12]. Datasets like FERET, XM2VTS DB, and LFW offer diverse sets of images under controlled laboratory conditions[34][36]. YouTube, SFC, and PaSC datasets contain variable conditions and a mix of images and videos. CelebFaces and CASIA WebFace present large datasets with variable conditions and resolutions. MegaFace and MS-Celeb-1M offer extensive datasets with a vast number of identities. VGGFace2, PIPA, and CFP datasets also contribute to the diversity of available data for face recognition research[39].

Table 1. Comparisons of facial detection dataset[1]

Dataset	No. of Images/Videos	No. of Ids	Resolution
MS-Celeb-1M [22]	8,556,240	98,892	300×300
VGGFace2 [23]	3.4M	9,100 +	Variable
PIPA [24]	64,18	2,456	Variable
CFP	7,100	510	Variable
LFW [10]	14,233	5,849	250×250
FERET [8]	15,051	274	512×768
XM2VTSDB [9]	2,460	285	720×576
YouTube [1][11]	3,525 vids	1,695	Variable
CMU Multi-PIE [12]	760,001	347	High-Res
SFC [13]	4.5M	4,130	Images
CAS-PEAL [14]	98,594	1,140	640×480
COX Face [15]	1,100 + 1,000 vids	1,100	Unknown
MegaFace [21]	4.9M	682,057	Variable
PaSC [16]	9,476 + 2,802 vids	283	Unknown
CelebFaces [17]	212,599	11,177	178×218
CASIA WebFace [19]	484,414	11,575	250×250
IJB-A [20]	5,812 + 2,085 vids	510	Variable

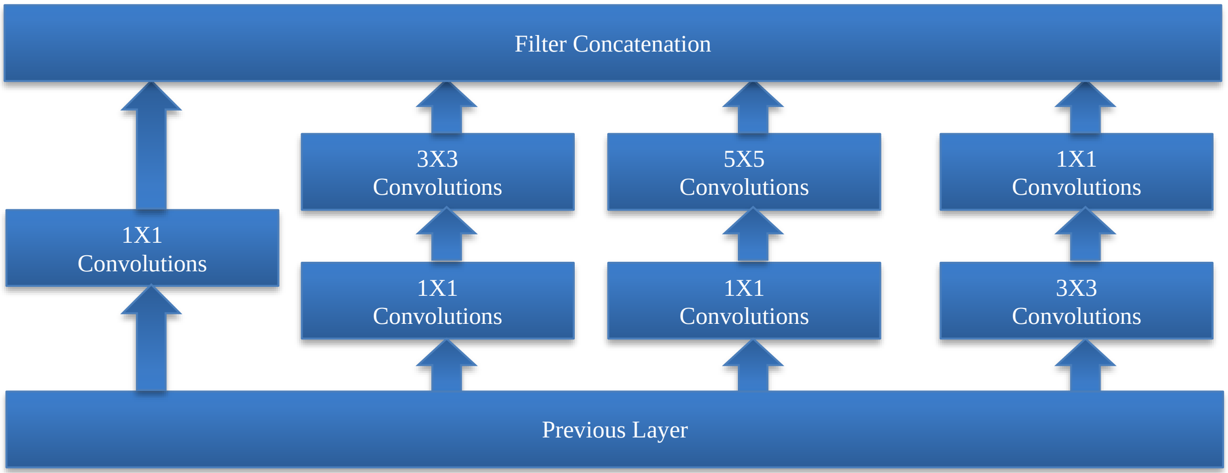

Fig. 4 The inception module in its normal configuration

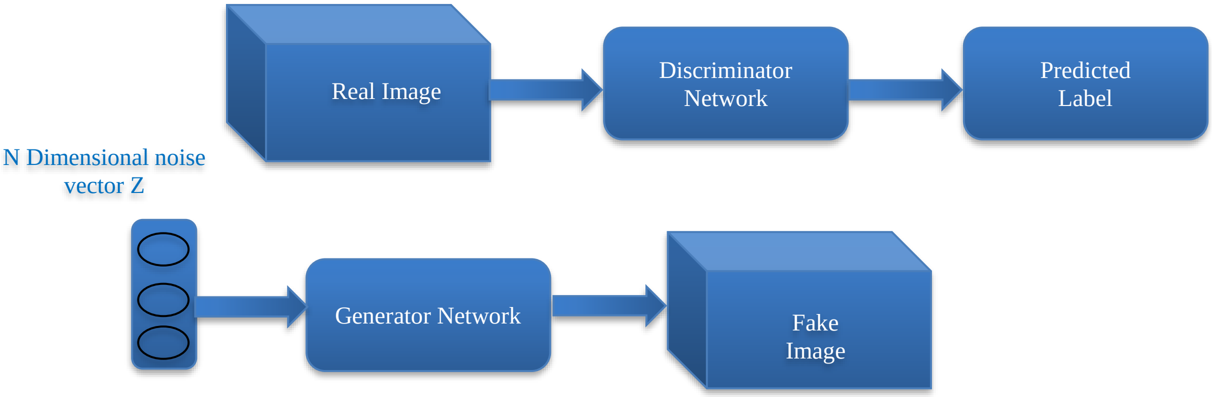

Fig. 5 Structure of conventionally generated adversarial networks

The standard version of the inception module is a building block in Convolutional Neural Networks (CNNs) designed to capture information at multiple scales within an image. It comprises parallel convolutional layers of different filter sizes and pooling operations[3]. This setup allows the network to learn features at various resolutions simultaneously, Improving its capacity to extract different and useful characteristics from input data[10]. The resulting feature maps from each parallel branch are then concatenated and passed on to subsequent layers for further processing. This architecture enables efficient utilization of computational resources while improving the network's performance in tasks like image classification and object detection.

The discriminators and the generator are the two main components of a conventional Adversarial Generation Network (GAN)[1]. The generator's purpose is to produce artificial data from an arrangement that closely resembles actual information[7][9]. It creates samples repeatedly, using random noise as the initial input. In contrast, the discriminator serves as a critic and makes an effort to discern between created and actual samples. These two networks compete with one another during the training process: the discriminator seeks to become more skilled at identifying bogus data, while the generator seeks to create more realistic data[15]. Both networks progressively get better through this adversarial training process until the generator produces data that the discriminator cannot tell apart from actual data. This framework is commonly utilized for tasks like picture production, data augmentation, and image-to-image translation.

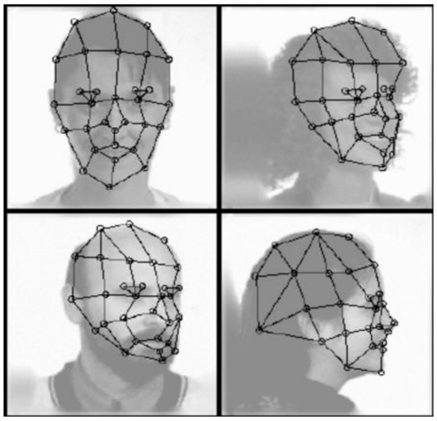

Fig. 6 Elastic bunch graph matched to faces

Wiskott et al. presented Elastic Bunch Graph Matching (EBGM) as a graph-matching technique. It creates a topological structure that connects every node to a specific collection of Gabor wavelets, which are renowned for their resilience to distortion, scaling, and variations in illumination[2]. EBGM uses a deformable matching method, which allows nodes to change their scale and location in response to alterations in face appearance, in contrast to standard graph matching. Even though EBGM performs competitively in face recognition tasks and is resilient to appearance changes, it has a high computational cost. It only uses feature point locations, ignoring additional picture information. Furthermore, the graph arrangement for initial faces is a challenge when done by hand; however, Campadelli and Lanzarotti's advances employ parametric models to solve this problem. An alternate method by Biswas et al. concatenates these characteristics for face representation in face recognition tasks by using SIFT features to characterize each landmark.

4. Result and Comparison

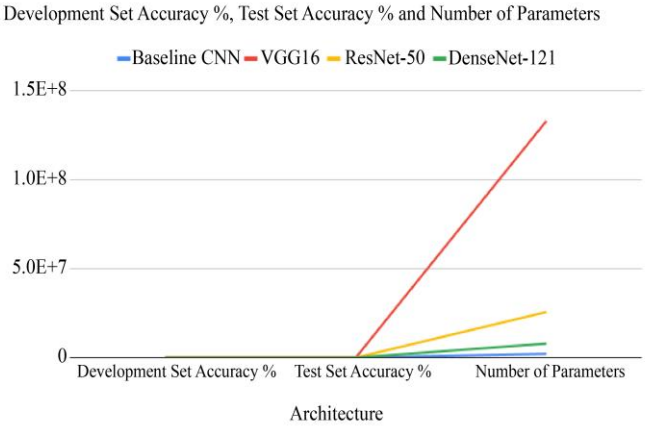

Fig. 7 An analysis of tested state-of-the-art models' classification rates for recognizing [1]

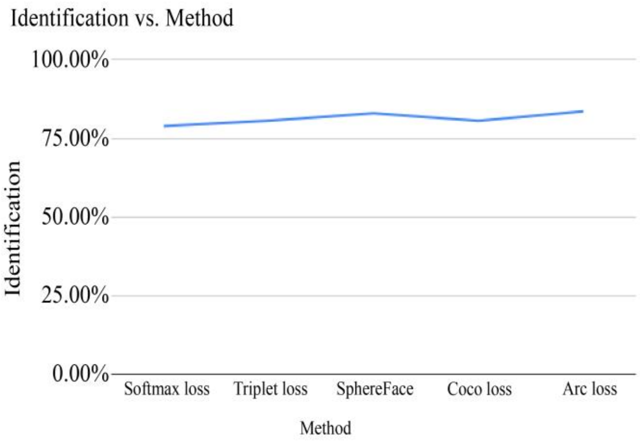

Fig. 8 Outcomes of the MegaFace challenge are compared for various functions of loss

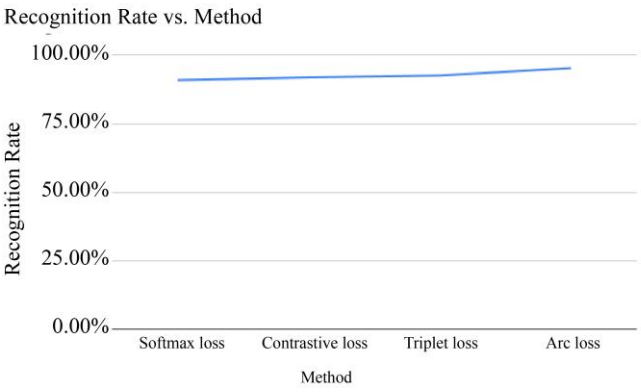

Fig. 9 Comparison of certain loss functions' categorization rates for recognition

5. Quantitative Results Comparison

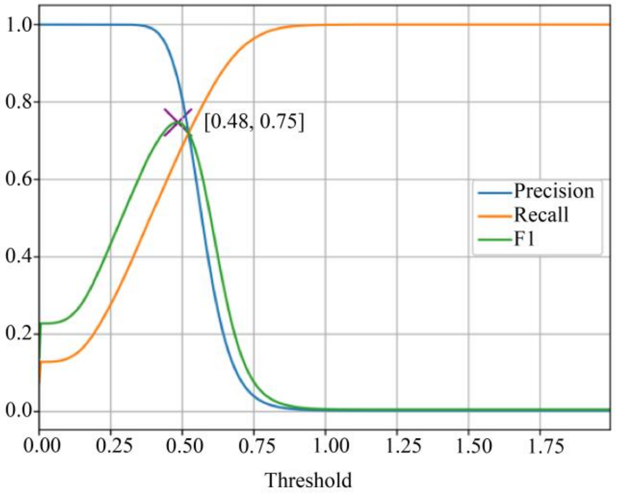

Fig. 10 F1 Score, Precision, and Recall for the color-FERET dataset. A purple cross denotes the optimal F1 score

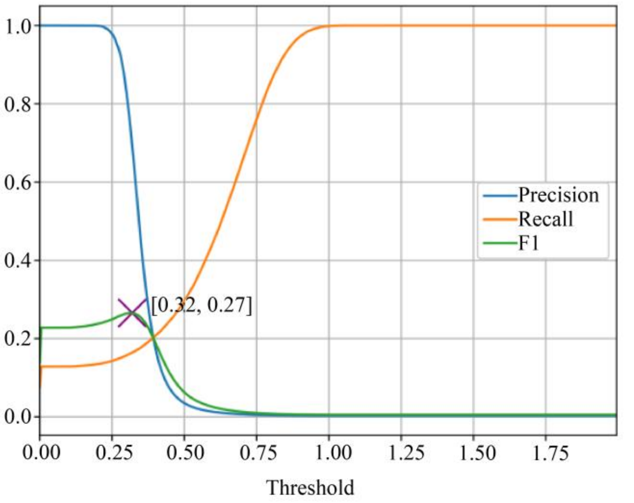

Fig. 11 Precision, Recall, and F1 Score for the Pix2Pix method-translated dataset. A purple cross denotes the optimal F1 score

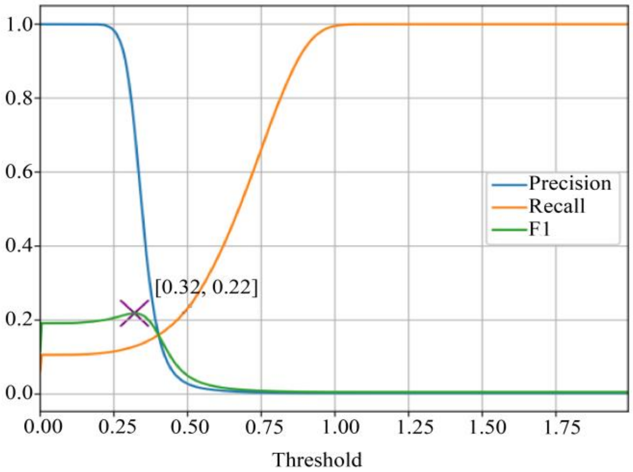

Fig. 12 The dataset's F1 Score, Precision, and Recall were converted using the UNIT technique. A purple cross denotes the optimal F1 score

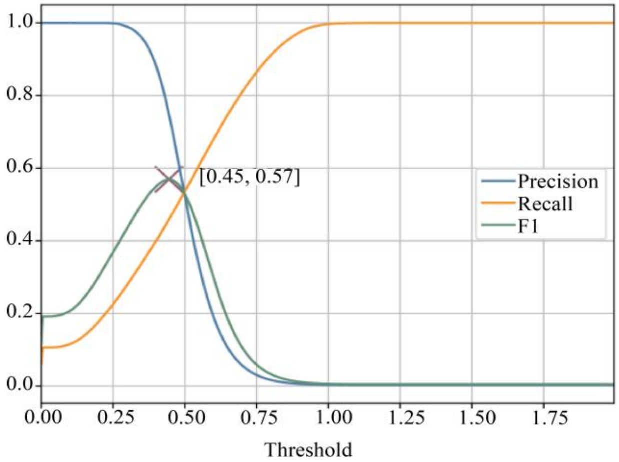

Fig. 13 For dataset reconstruction using the X-Bridge approach, precision, recall, and F1 score are calculated. A purple cross denotes the optimal F1 score.

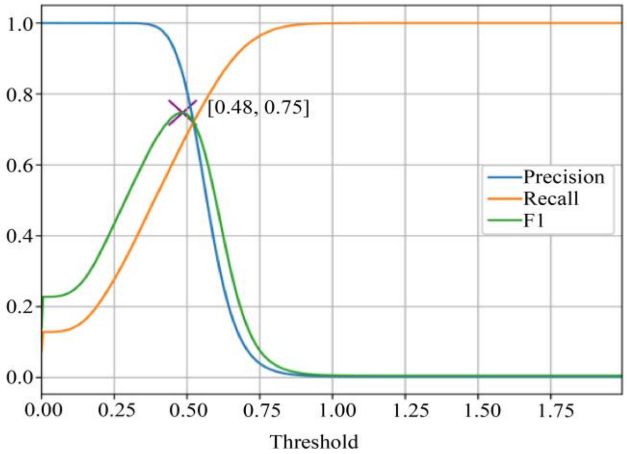

Fig. 14 For datasets translated using the X-Bridge approach, precision, recall, and F1 score were obtained. A purple cross indicates the optimal F1 score

Table 2. Method vs F1 score

Method	Dataset	F1 Score	Specific Threshold	Comments
ArcFace Classifier	Color-FERET	0.75	-	Achieved F1 Score: 0.75. The classifier yielded an improved F1 Score of 0.80 for a specific distance threshold.
Pix2pix Method	Translated	-	0.27	Significant performance drop observed with ArcFace classifier on the translated dataset. Challenges attributed to the classifier's training on photos rather than sketches.
UNIT Method	Translated	-	-	The UNIT method yielded inferior results compared to Pix2pix. Despite being robust enough to handle pose changes, the translated sketch quality remains lower.
X-Bridge Method	Translated	0.57	-	Outperformed other methods significantly, achieving an F1 Score of 0.57 for a specific threshold. A combination of accurate translation and robustness in pose and expression contributes to a performance boost.

ArcFace Classifier achieved an F1 Score of 0.75, improving to 0.80 with a specific threshold. Pix2pix saw a drop to the F1 Score of 0.27 due to photo-based training.

UNIT's results were inferior to those of Pix2pix, while X-Bridge outperformed others with a 0.57 F1 Score thanks to accurate translation and robustness.

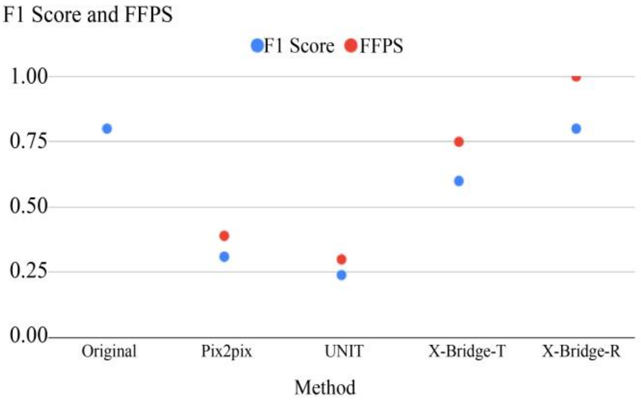

Fig. 15 Comparison of the tested Cross-modal bridges.

6. Conclusion

The present research presents a unique heterogeneous Image of Faces program that uses X-Bridge, a synthesis-based cross-modal bridge[7]. By converting pictures between two modalities while maintaining important face traits, the system compensates for differences between the modalities. Before producing translated pictures based on the acquired latent coding, X-Bridge encodes input images into a common latent space. A DenseNet-based facial extraction algorithm that has been taught on the Casia-WebFace dataset utilizing Arc loss for greater separation of classes then analyzes the converted photographs[1]. Comparisons with traditional Softmax show significant improvement, particularly in open-set classification. The system, which consists of the feature extractor and cross-modal bridge, performs better than previous approaches and is assessed using a brand-new measure known as the Facial Feature Preservation Score (FFPS). The research analyzes face recognition datasets, neural network architectures, and loss functions to identify the most suitable options. It also explores existing methods for cross-modal bridge applications, focusing on generative adversarial networks. A novel methodology known as X-Bridge is introduced that expands on prior techniques and offers state-of-the-art qualitative and quantitative outcomes[11].

References

[1] Ishaan Gulrajani et al., "Improved Training of Wasserstein Gans," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach California, USA, pp. 5769-5779, 2017. [Google Scholar] [Publisher Link]

[2] Anders Boesen Lindbo Larsen et al., "Autoencoding Beyond Pixels Using a Learned Similarity Metric," *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, pp. 1558-1566, 2016. [Google Scholar] [Publisher Link]

[3] Ivan Gruber, "Generating Facial Images Using Vaegan," *Student Scientific Conference*, pp. 38-39, 2018. [Google Scholar] [Publisher Link]

[4] Mehdi Mirza, and Simon Osindero, "Conditional Generative Adversarial Nets," *arXiv*, pp. 1-7, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[5] Diederik P. Kingma, and Max Welling, "Auto-Encoding Variational Bayes," *arXiv*, pp. 1-14, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[6] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, pp. 1735-1742, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yi Sun et al., "Deep Learning Face Representation by Joint Identification-Verification," *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal Canada, vol. 2, pp. 1988-1996, 2014. [Google Scholar] [Publisher Link]

[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 815-823, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[9] Yandong Wen et al., "A Discriminative Feature Learning Approach for Deep Face Recognition," *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, pp. 499-515, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[10] Yu Liu, Hongyang Li, and Xiaogang Wang, "Learning Deep Features via Congenerous Cosine Loss for Person Recognition," *arXiv*, pp. 1-7, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] Jiankang Deng et al., "Arcface: Additive Angular Margin Loss for Deep Face Recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 4685-4694, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Rajeev Ranjan, Carlos D. Castillo, and Rama Chellappa, "L2-Constrained Softmax Loss for Discriminative Face Verification," *arXiv*, pp. 1-10, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[13] Feng Wang et al., “Normface: L_2 Hypersphere Embedding for Face Verification,” *Proceedings of the 25th ACM International Conference on Multimedia*, Mountain View California, USA, pp. 1041-1049, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[14] Hao Wang et al., “Cosface: Large Margin Cosine Loss for Deep Face Recognition,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 5265-5274, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[15] Feng Wang et al., “Additive Margin Softmax for Face Verification,” *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926-930, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] A Beginner’s Guide to Generative Adversarial Networks (GANS), Skymind. [Online]. Available: https://skymind.com/wiki/generative-adversarial-network-gan

[17] Cs231n Convolution Neural Networks for Visual Recognition. [Online]. Available: http://cs231n.github.io/

[18] "Neuronov´e s´ıtˇe." [Online]. Available: http://www.kky.zcu.cz/cs/courses/neu

[19] T. Kohonen, M. Schroeder, and T. Huang, *Self-Organizing Maps*, 3rd ed., Springer-Verlag New York, Secaucus, NJ, USA, 2001. [Publisher Link]

[20] Lior Wolf, Tal Hassner, and Itay Maoz, “Face Recognition in Unconstrained Videos with Matched Background Similarity,” *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, pp. 529-534, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[21] Ralph Gross et al., “Multi-Pie,” *Image and Vision Computing*, vol. 28, no. 5, pp. 807-813, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[22] Yaniv Taigman et al., “Deepface: Closing the Gap to Human-Level Performance in Face Verification,” *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1701-1708, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[23] Wen Gao et al., “The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations.” *IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems and Humans*, vol. 38, no. 1, pp. 149-161, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[24] Zhiwu Huang et al., “A Benchmark and Comparative Study of Video-Based Face Recognition on Cox Face Database,” *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5967-5981, 2015. [CrossRef] [Google Scholar] [Publisher Link]