International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(17s) 2026 pp. 741-751
https://doi.org/10.70917/ijcisim-2026-4796
Article

Self-Supervised Capsule Network for Robust Face Recognition

K. Minney Prisilla[1], N. Jayashri[2]

[1]Dr. M.G.R . Educational and Research Institute, Chennai.
minprisy1812@gmail.com
[2] Faculty of Computer Applications, Dr. M.G.R. Educational and Research Institute, Chennai
jayashrichandrasekar@yahoo.co.in

Abstract: Face recognition becomes one of the most adopted biometric technics due to its applications in intelligent surveillance, access control, border security, digital authentication, criminal investigation and human computer interaction. The development of Deep convolutional neural networks (CNNs) significantly improved accuracy of recognition even in unconstrained environments such as pose variations, illumination changes, facial expressions, occlusions and low-resolution images. Conventional CNNs mainly focus on local spatial features and so it has limited ability to preserve hierarchical association between facial components. Capsule Networks (CapsNets) overcome this by representing visual features as vector capsules with existence and geometric properties of objects. The self supervised learning of CapsNets enables feature learning from unlabelled images. This article presents a Self supervised Capsule Network (SS-CapsNet) integrates convolutional feature extraction, capsule based learning, dynamic routing and contrastive self supervised learning into a combined framework for robust face recognition. This approach simultaneously learns discriminative identity from large scale unlabelled dataset while preserves facial geometry. The SS-CapsNet provides improved robustness against pose variations, illumination changes, facial occlusions and image degradation.
Keywords: Face Recognition; Capsule Network; Self supervised Learning; Contrastive Learning; Dynamic Routing; Deep Learning; Computer Vision

1. Introduction

Due to the widespread applications of security, surveillance, digital identity management, financial authentication, healthcare and smart city infrastructure, automated face recognition is considered as one of the influential research areas in computer vision. The availability of large facial image repositories accelerated the development of deep learning models for human face recognition with extraordinary performance [1]. These models are expected to operate in challenging real world conditions with variations of illumination, facial expression, aging, head pose, image blur, occlusion and background clutter [2][3].

Earlier face recognitions primarily trusted on handcrafted feature descriptors like Eigenfaces, Fisherfaces, Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG) and Scale Invariant Feature Transform (SIFT) [4]-[9]. Based on statistical appearance or local texture information, these descriptors extract manually designed characteristics from face [10]. However, they generated satisfactory recognition in controlled imaging conditions [11]. But handcrafted descriptors will not capture complex facial variations effectively from large scale datasets [12].

Instead of descriptors' manual design, CNNs automatically learn hierarchical representation of images directly from images through repeated convolution, nonlinear activation, normalization and pooling operations [13][14]. Preliminary convolutional layers detect important visual structures including edges, corners and texture patterns [15]. Intermediate layer converts these low level features into facial landmarks such as eyebrows, eyes, nose, lips and facial contours. Finally. deeper convolutional layer integrates these features into semantic identity representation for face recognition [16].

Despite of their remarkable achievements, CNN based architectures has certain limitations. Convolution kernels process only localized neighbourhoods of an image, restricting the ability to model spatial relationships between distant facial components. However, deeper network increases receptive fields to capture long range dependencies indirectly through successive convolution operations. Subsequently, conventional CNNs may not preserve geometric associations between facial landmarks in severe pose variation [17] [18].

In recent researches, advanced CNN architectures are proposed to overcome these limitations. ResNet introduces residual learning to enable training of very deep networks by implementing shortcut connections [19]. DenseNet is used to improve feature reuse through dense connectivity of layers [20]. MobileNetV2 using inverted residual bottlenecks and depthwise separable convolution to reduce computational complexity [21]. EfficientNet balances network depth, width and resolution through compound scaling [22]. These architectures represent features primarily as scalar activations and so improved recognition accuracy [23].

Capsule Networks provide an alternative representation by preserving hierarchical relationships of visual entities. Capsules produce multidimensional vectors instead of scalar outputs generated by conventional neurons. The length of these vectors represents the probability of object existence while their orientations encode pose, rotation, scale and deformation information. Dynamic routing enables lower level capsules to transmit information to higher level capsules and so it preserves spatial consistency among facial components. Accordingly, Capsule Networks demonstrate improved robustness in affine transformations [24][25].

Manual annotation of large facial image collections is expensive, and often impractical. Self supervised learning is considered as an alternative by enabling deep neural networks to learn feature representations from unlabelled data [26]. Instead of relying on class labels, self supervised learning generates supervisory signals from intrinsic relationship within the data. Contrastive learning methods such as SimCLR, MoCo, BYOL and DINO learn features by maximizing resemblance between different augmented views of same person while separating distinct images [27]-[30].

Integrating self supervised learning with Capsule Networks combines the complementary strengths of both paradigms. Self supervised representation learning extracts generalized facial representations from large scale unlabeled dataset. Capsule routing preserve hierarchical spatial relationship between facial landmarks. Therefore, the resulting embeddings become extremely discriminative and robust against pose variation, illumination changes, facial expressions, occlusion and image degradation. It extracts local visual patterns through convolutional layers, transforms them into vector capsules, performs routing-by-agreement to preserve hierarchical facial relationships and finally learns discriminative embeddings using self-supervised contrastive optimization.

2. Proposed Self-Supervised Capsule Network (SS-CapsNet)

The proposed Self Supervised Capsule Network (SS-CapsNet) is designed by integrating the strengths of CNN, capsule networks and self supervised learning into a united architecture to improve face recognition. CNNs learns local texture representations but fail to preserve spatial relationships of facial components. Capsule Networks address this issue by encoding facial properties as vectors to retain geometric data such as orientation, scale and pose. Self supervised learning is to learn highly discriminative representations from unlabelled facial images.

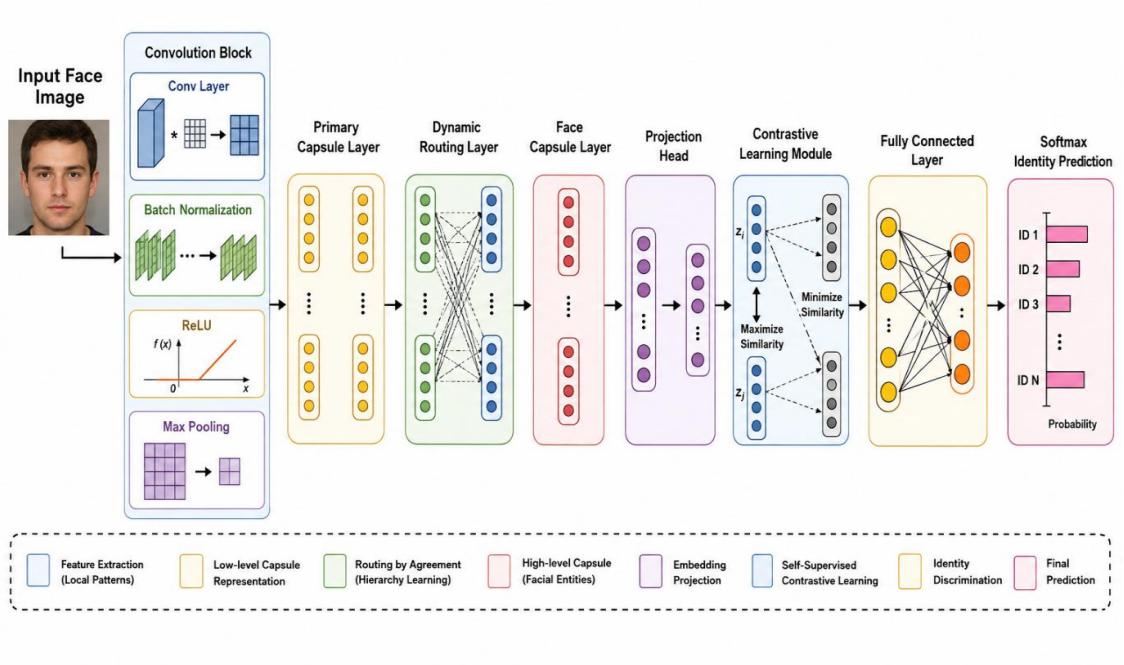

Fig. 1: The structure of proposed Self Supervised Capsule Network

2.1 Convolutional Feature Extraction

The convolution block makes the initial feature extraction by learning local visual characteristics. It consists of four sequential operations like convolution layer (Conv Layer), Batch Normalization, ReLU Activation and Max Pooling. The conv layer applies multiple filters on input image to detect local image patterns. Each filter is for specific structures like edges, corners, texture variations and fine facial details. Batch normalization standardizes the feature distributions. It makes learning process more efficient and reduces internal covariate shift. Rectified Linear Unit (ReLU) presents non-linearity into the network and so it learns complex relationships among facial features by suppresses negative responses. Max pooling reduces the spatial dimensions while retaining the most important responses.

The captured RGB face image is initially passed through multiple convolution layers,

$$F(j) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} I(+mj + n)W(n) + b$$

where,

I= Input image

W= Convolution kernel

b= Bias

k= Kernel size

$F(j)$= Output feature map

After convolution,

ReLU activation is applied

$$ReLU(x) = max(x)$$

The feature maps are then downsampled using max pooling

$$P(j) = \underset{(n)\in\Omega}{max} F(n)$$

where,

Ω denotes the pooling region.

The output of the convolution block represents local facial structures like eyes, nose, eyebrows, lips, skin texture and facial contours.

2.2 Primary Capsule Layer

Primary Capsule Layer converts convolutional feature maps into vector characters called capsules. Each capsule comprises multiple numerical values to define geometric properties of a facial component. Suppose the convolution layer generates N feature vectors, each vector is transformed into a capsule u_i using,

$$\hat{u}_{j|i} = W_{ij} u_i$$

where,

u_i= Input capsule

W_{ij}= Learnable transformation matrix

$\hat{u}_{j|i}$= Prediction vector

Instead of a single activation value, each capsule represents position, orientation, scale, texture and facial geometry.

2.3 Dynamic Routing Layer

Dynamic routing replaces conventional pooling operations while maintaining the geometric consistency of facial structures. It identifies the lower-level capsules that can contribute to higher-level capsules.

Routing coefficients are computed as,

$$c_{ij} = \frac{exp(b_{ij})}{\sum_k \quad exp(b_{ik})}$$

where,

b_{ij} represents routing logits.

Therefore, the higher level capsule input becomes,

$$s_j = \sum_i \quad c_{ij} \hat{u}_{j|i}$$

The capsule output is determined by using the squashing function,

$$v_j = \frac{\| s_j \|^2}{1+\| s_j \|^2} \frac{s_j}{\| s_j \|}$$

This squashing makes short vectors to zero near value and long vectors to unit length. Therefore, vector length represents the existence probability of facial entity and vector orientation stores pose information.

2.4 Face Capsule Layer

The Face Capsule Layer consolidate all information that received from routing stage to create compact representation of the face. Each capsule corresponds to higher level entity rather than individual local feature of the face.

2.5 Projection Head

The projection head transform capsule results into a latent embedding space for self supervised learning. It consists of one or more fully connected layers to reduce dimensionality of capsule representations. In self supervised pre-training, the projection head generates compact feature embeddings through contrastive learning. After that, it will be removed or replaced by an identity classification layer.

2.6 Capsule Margin Loss

After the self supervised pre-training, capsule margin loss is using for the identity classification.

The loss function is,

$$L_k = T_k max(m^+ - || v_k ||)^2 + \lambda(1 - T_k) max(|| v_k || - m^-)^2$$

where,

T_k= Ground truth label

m^+= Positive margin

m^-= Negative margin

λ= Down-weighting coefficient

Here, the correct identity capsules generate large vector lengths and incorrect capsules

generate small vector lengths.

2.7 Contrastive Learning

In contrastive learning, it performs self supervised learning without manual identity labels. Instead of class labels, self supervised learning creates two augmented versions of same face by using transformations such as random cropping, horizontal flipping, brightness adjustment, Gaussian blur and colour perturbation.

Let x be the original image. Therefore, the two augmented samples are,

$$x_i = T_1(x) x_j = T_2(x)$$

where,

T_1 and T_2 represent random augmentations such as random crop, horizontal flip, colour jitter, Gaussian blur. Both the images are processed through the same capsule encoder. Therefore, the obtained embeddings are,

$$z_i = f(x_i) z_j = f\left(x_j\right)$$

Positive samples have similar embeddings.

2.8 Contrastive Loss Function

The proposed model employs the InfoNCE loss.

$$L_{SSL} = -\log \log \frac{exp\ exp\left(\frac{sim(z_j)}{\tau}\right)}{\left(\frac{sim(z_k)}{\tau}\right)}$$

where,

sim= Cosine similarity

τ= Temperature parameter

N= Batch size

It minimizes the distance between positive image pairs and maximizes separation from

negative image pairs.

2.9. Fully Connected Layer

The fully connected layer receives learned facial embeddings from capsule network and transforms it into identity specific feature vectors. This stage combines geometric information of compact representation for the classification. It is the final feature transformation before the prediction of identity.

2.10. Softmax Identity Prediction

The Softmax layer makes final identity classification by converting the outputs of fully connected layer into a probability distribution. Output probability characterizes the likelihood of the input face that belongs to a specified individual. The identity of the highest probability is designated as recognition result. In addition to this, Softmax probability provides confidence score for the recognition decision.

2.11 Reconstruction Loss

The model reconstructs the original image to preserve facial information. The reconstruction loss is,

$$L_{rec} = || x - \hat{x} ||^2$$

where,

x= Original image

$\hat{x}$= Reconstructed image

The reconstruction decoder forces capsules to encode meaningful facial information.

The total optimization combines all three losses,

$$L = L_{SSL} + \alpha L_{margin} + \beta L_{rec}$$

where,

αcontrols supervised learning

βcontrols reconstruction learning

This joint optimization learns about discriminative embeddings, spatial relationships, geometric consistency and robust identity representations simultaneously.

The proposed system is trained in two phases with WIDER FACE dataset where initially large set of unlabelled facial images are used for self supervised pre-training. Two augmented views of each image are generated and passed through convolutional and capsule encoder. The parameters are optimized in contrastive loss to maximize embeddings of the same image while differentiating dissimilar images. In second phase, the pre-trained encoder is tuned with labelled dataset. The Adam optimizer with cosine learning rate continuously update parameters to minimize loss function and ensures stable convergence. These two stage learning produce highly discriminative facial embeddings.

3. Performance Evaluation Metrics

The efficiency of the SS-CapsNet is calculated by using face recognition performance metrics such as accuracy, precision, recall, F1-score, parameters and receiver operating characteristic (ROC).

Table 1. Comparative Performance Evaluation of SS-CapsNet

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Parameters (Million)**
CNN	94.8	94.4	94.1	94.2	5.23
ResNet-50	96.3	96.2	95.8	96.0	25.56
DenseNet-121	96.9	96.8	96.6	96.8	7.98
MobileNetV2	95.7	95.4	95.3	95.3	3.51
EfficientNet-B0	97.5	97.2	97.1	97.2	5.27
Vision Transformer	97.8	97.6	97.5	97.6	86.0
Swin Transformer	98.2	98.0	97.9	97.9	28.2
Proposed SS-CapsNet	98.8	98.6	98.5	98.5	10.4

Table 1 shows comparative performance of different deep learning models in face recognition. In CNN-based architectures, EfficientNet-B0 achieved the highest accuracy of 97.5% with 5.29 million parameters. It indicate an effective balance between recognition rate and model complexity. ResNet-50 and DenseNet-121 also provided high accuracy, precision, recall and F1-score. But, ResNet-50 required a larger number of parameters. MobileNetV2 demonstrates competitive performance with the smallest model size, making it suitable for resource constrained applications. Transformer based models like Swin Transformer improved recognition by effectively capturing global contextual information. However, the proposed SS-CapsNet outperforms all other models and achieved the highest accuracy value (98.8%), precision (98.6%), recall (98.5%) and F1-score (98.5%) with reasonable number of parameters as 10.4 million**.** These results shows that the hybridization of self supervised learning with capsule based feature representation improves recognition accuracy while maintaining computational efficiency.

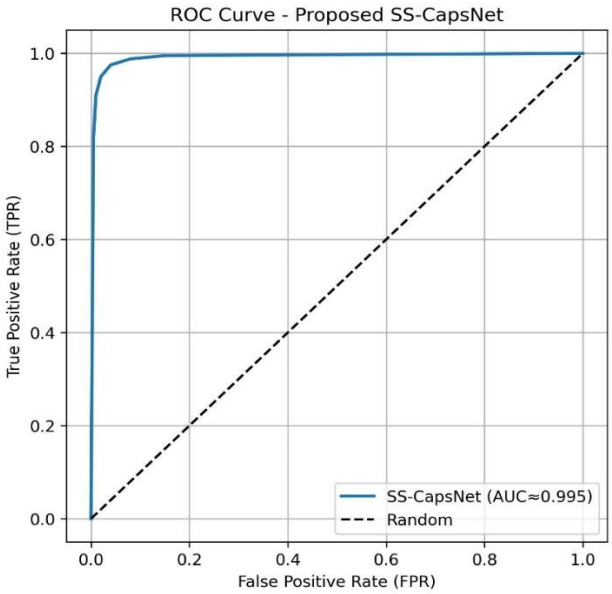

Fig. 2: ROC curve of Self Supervised Capsule Network

The ROC curve illustrates the classification performance of the proposed model. The curve lies close to the upper left corner indicates high true positive rate. The estimated Area Under the Curve of 0.995 demonstrates the model's t ability to distinguish genuine and non-genuine images. This result indicates that the model is well suitable with minimal classification errors.

4. Robustness Analysis

The robustness values are evaluated for each model separately under different imaging conditions. A subset of 650 images with specific variation like different head poses and varying illumination levels are created. The average robustness is computed by averaging the recognition accuracies of the models.

$$Average\ Robustness = \frac{\sum_{i=1}^{n} \quad R_i}{n}$$

where,

R_i= recognition accuracy for the i^{th}challenging condition,

n= total number of evaluated conditions (here, $n = 5$)

Table 2. Quantitative Robustness Analysis of Face Recognition Models in percentage

Challenge	**CNN**	**ResNet-50**	**EfficientNet-B0**	**Swin Transformer**	**Proposed SS-CapsNet**
Pose Variation	88.5	93.2	94.1	96.3	98.1
Illumination Change	89.1	94.4	96.2	96.8	98.4
Facial Occlusion	87.4	92.8	93.9	96.1	97.9
Low Resolution	86.9	89.8	92.5	93.7	97.3

Expression Variation	92.6	94.8	95.4	97.0	98.6
Average Robustness (%)	88.9	93.0	94.4	96.0	98.1

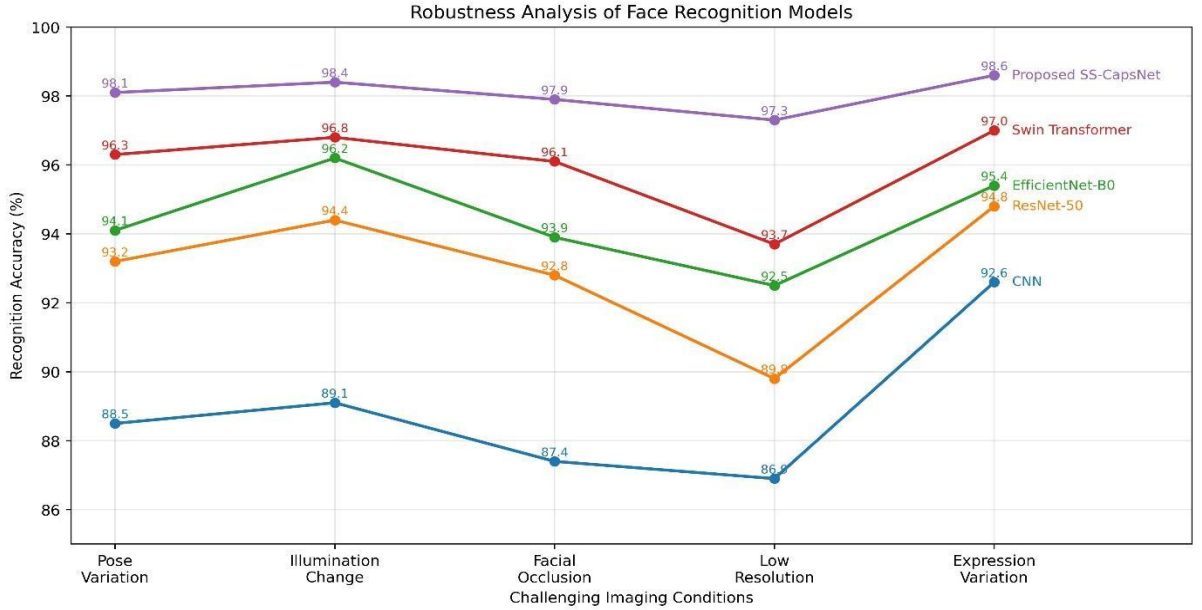

Fig. 3: Robustness Analysis of Face Recognition Models

The robustness analysis shows that recognition performance decreases for all models when facial images are affected by pose changes, illumination variation, occlusion, low resolution, and expression changes. Conventional CNN exhibits the largest performance degradation, particularly for low-resolution and occluded images. ResNet-50 and EfficientNet-B0 improve robustness through deeper feature learning and efficient feature extraction, while Swin Transformer further benefits from global contextual modeling. The proposed SS-CapsNet consistently achieves the highest recognition accuracy across all challenging conditions, with an average robustness of 98.1%. This improvement can be attributed to the capsule-based representation, which preserves spatial relationships among facial components, and the self-supervised learning strategy, which enhances feature generalization. Consequently, the proposed model provides more reliable face recognition in unconstrained real-world environments than the compared architectures.

5. Discussion

The experimental investigation demonstrates that the proposed Self Supervised Capsule Network successfully combines the complementary strengths of convolutional feature extraction, capsule representation learning and self supervised contrastive optimization. Conventional CNNs primarily focus on local texture information. but the capsule vectors encode both the probability and geometric properties of facial components.

The incorporation of self supervised learning substantially reduces the dependency on manually labelled dataset. In pretraining, the model learns generalized facial representations by maximizing agreement between different augmented views of the same face. These learned embeddings provide an excellent initialization for supervised fine tuning, leading to faster convergence and improved recognition performance. Dynamic routing plays an essential role in preserving spatial consistency among facial components. This system allows the network to maintain consistent identity representations despite changes in viewpoint, facial expression, illumination and occlusion.

6. Conclusion

This research presented a Self supervised Capsule Network (SS-CapsNet) for robust face recognition. This model integrates convolutional feature extraction, capsule representation learning, dynamic routing and self-supervised contrastive optimization into a single framework. It preserves geometric relationships among facial components by representing them as multidimensional capsules rather than scalar activations. Self supervised pretraining enables effective utilization of large scale unlabelled dataset and reduces the dependency on manual annotation. The experimental evaluation demonstrates that SS-CapsNet achieves superior recognition accuracy and robustness under challenging imaging conditions including pose variation, illumination changes, facial expression, occlusion and low resolution images. Comparative analysis with CNN, ResNet, DenseNet, MobileNetV2, EfficientNet, Vision Transformer and Swin Transformer confirms that the proposed framework effectively balances recognition performance, computational efficiency and representation capability. These characteristics makes SS-CapsNet a competitive model for next generation intelligent surveillance and face recognition systems.

REFERENCES

1. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
2. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
3. M. Joseph and K. Elleithy, "Beyond frontal face recognition," *IEEE Access*, vol. 11, pp. 26850–26861, 2023, doi: 10.1109/ACCESS.2023.3258444.
4. H.-T. Ho, L. V. Nguyen, T. H. T. Le, and O.-J. Lee, "Face detection using Eigenfaces: A comprehensive review," *IEEE Access*, vol. 12, pp. 118406–118426, 2024, doi: 10.1109/ACCESS.2024.3435964.
5. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997, doi: 10.1109/34.598228.
6. N. Arora, I. Budhiraja, D. Garg, S. Garg, B. J. Choi, and M. S. Hossain, "Revolutionizing facial image retrieval: Multi-block and mean based local binary patterns with sign and magnitude analysis," *Alexandria Engineering Journal*, vol. 116, pp. 601–608, Mar. 2025, doi: 10.1016/j.aej.2024.12.003.
7. S. S. Greeshan, S. Maloji, and K. Mannepalli, "Facial recognition using HOG algorithm and SVM classifier," *Journal of Information Systems Engineering and Management*, vol. 10, no. 42s, pp. 447–454, 2025, doi: 10.52783/jisem.v10i42s.7904.
8. H. Li, L. Wang, T. Zhao, and W. Zhao, "Local-peak scale-invariant feature transform for fast and random image stitching," *Sensors*, vol. 24, no. 17, Art. no. 5759, 2024, doi: 10.3390/s24175759.
9. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
10. A. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey," *Sensors*, vol. 20, no. 2, Art. no. 342, 2020, doi: 10.3390/s20020342.
11. A. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electronics*, vol. 9, no. 9, Art. no. 1448, 2020, doi: 10.3390/electronics9091448.
12. M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021, doi: 10.1016/j.neucom.2020.10.081.
13. H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei, "The elements of end-to-end deep face recognition: A survey of recent advances," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 126–141, 2022, doi: 10.1109/TIFS.2021.3139536.
14. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, 2014, pp. 1701–1708, doi: 10.1109/CVPR.2014.220.
15. M. Wang, Y. Zhang, and W. Deng, "Meta face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 11273–11282.
16. J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022, doi: 10.1109/TPAMI.2021.3087709.
17. O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and F. Khelifi, "Pose-invariant face recognition with multitask cascade networks," *Neural Computing and Applications*, vol. 34, no. 8, pp. 6039–6052, Apr. 2022, doi: 10.1007/s00521-021-06690-4.
18. X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognition*, vol. 42, no. 11, pp. 2876–2896, Nov. 2009, doi: 10.1016/j.patcog.2009.04.017.
19. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

20. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.
21. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.
22. M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.
23. S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
24. J. Gu and V. Tresp, “Improving the robustness of capsule networks to image affine transformations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, 2020, pp. 7283–7291, doi: 10.1109/CVPR42600.2020.00731.
25. C. Pan and S. Velipasalar, “PT-CapsNet: A novel prediction-tuning capsule network suitable for deeper architectures,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 11996–12005, doi: 10.1109/ICCV48922.2021.01180.
26. T. Uelwer, J. Robine, S. S. Wagner, M. Höftmann, E. Upschulte, S. Konietzny, M. Behrendt, and S. Harmeling, “A survey on self-supervised methods for visual representation learning,” *Machine Learning*, vol. 114, Art. no. 111, 2025, doi: 10.1007/s10994-024-06708-7.
27. J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow Twins: Self-supervised learning via redundancy reduction,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 12310–12320.
28. M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 9650–9660, doi: 10.1109/ICCV48922.2021.00951.
29. X. Chen and K. He, “Exploring simple Siamese representation learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 15750–15758, doi: 10.1109/CVPR46437.2021.01549.
30. M. Assran, M. Caron, I. Misra, P. Bojanowski, A. Joulin, N. Ballas, and M. Rabbat, “Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, 2021, pp. 8443–8452, doi: 10.1109/ICCV48922.2021.00833.