Robust Facial Feature Extraction Using Deep Learning Under Varying Image Conditions

Aldy Agustian, Muhammad Syahputra Novelan, Solly Aryza

Abstract
Facial recognition has become one of the most prominent applications of computer vision, driven by advances in deep learning techniques. The performance of facial recognition systems is strongly influenced by the quality of feature representations and the model's ability to generalize across different image conditions. In real-world applications, facial images are frequently affected by variations in illumination, facial pose, expression, image resolution, and occlusion, which may influence feature consistency and recognition accuracy. This study aims to investigate facial feature extraction and evaluate the generalization capability of deep learning models under varying image conditions. The proposed approach employs a Convolutional Neural Network (CNN) as a feature extractor, with pretrained architectures and embedding-based representation learning. The evaluation is designed around illumination variations, pose changes, resolution degradation, expression variations, and partial facial occlusion. Model performance is assessed using accuracy, precision, recall, F1-score, cosine similarity, and Receiver Operating Characteristic–Area Under the Curve (ROC-AUC), subject to an appropriate face-verification protocol. Embedding distribution analysis is also used to investigate the stability of facial representations across image conditions. The study is expected to clarify the robustness of facial feature representations, the sensitivity of deep learning models to image variations, and the factors affecting model generalization. The findings may contribute to the development of more reliable facial recognition systems for unconstrained environments..

***Keywords**:*
Deep Learning; Facial Feature Extraction; Face Recognition; Convolutional Neural Network; Feature Embedding; Model Generalization; Computer Vision.

Aldy Agustian
Universitaas Pembangunan Panca Budi, Indonesia
e-mail: aldy.agustian720@gmail.com

Muhammad Syahputra Novelan, Solly Aryza
e-mail: putranovelan@dosen.pancabudi.ac.id , sollyaryzalubis@pancabudi.ac.id
3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction

The rapid development of artificial intelligence has significantly expanded the application of deep learning across various fields, particularly digital image processing and computer vision. One of its major applications is facial recognition, a technology that enables computer systems to identify or verify individuals based on facial characteristics.

Conventional facial recognition approaches generally rely on manually engineered geometric and texture-based features. However, these methods may encounter limitations when processing images captured under varying conditions. The emergence of deep learning, particularly Convolutional Neural Networks (CNNs), has introduced a more flexible approach by enabling automatic feature extraction directly from image data.

In deep learning-based facial recognition systems, feature representation plays a fundamental role because it determines how facial characteristics are transformed into numerical representations, commonly referred to as feature embeddings. Ideally, these representations preserve identity-related information while reducing the influence of visual variations that are unrelated to identity.

Although deep learning models have demonstrated strong performance on benchmark datasets, high performance on a particular test set does not necessarily indicate strong generalization. Challenges arise when models are applied to facial images that differ from the training distribution, including images with low illumination, non-frontal poses, different facial expressions, low resolution, or partial occlusion.

Embedding-based methods such as FaceNet and ArcFace have advanced facial representation learning. FaceNet uses triplet-based learning to structure the embedding space, while ArcFace introduces an additive angular margin objective to improve discrimination between identities. Nevertheless, robustness under uncontrolled image conditions remains an important issue that requires systematic evaluation.

This study therefore investigates facial feature extraction and model generalization under varying image conditions. In addition to recognition performance, the study considers the consistency of feature embeddings when facial images undergo controlled visual transformations.

Literature Review

2.1 Deep Learning for Face Recognition

Deep learning uses multilayer neural networks to learn hierarchical representations from data. In facial recognition, CNNs extract visual information through convolutional operations, nonlinear activation functions, and spatial aggregation. Early layers typically learn low-level patterns such as edges and textures, while deeper layers learn more complex facial structures and identity-related characteristics.

2.2 Facial Feature Representation

Facial feature representation transforms a facial image into a numerical vector that encodes identity-related information. A general feature extraction function can be expressed as:

$$F(x) = f\theta(x)$$

where x is the input facial image, fθ is the feature extraction function parameterized by the model, and F(x) is the resulting feature representation. Images of the same individual are expected to produce embeddings that are relatively close in the feature space, whereas embeddings from different individuals should be more separable.

2.3 Convolutional Neural Network Architecture

CNNs are widely used in image processing because they can learn spatial patterns from image data. A typical recognition pipeline may include a convolutional backbone, nonlinear activation, feature aggregation, and an embedding layer. The final embedding is used for similarity measurement or classification.

1. Convolution layer: extracts local visual patterns and facial characteristics.
2. Activation function: introduces nonlinear transformations.

3. Pooling or feature aggregation: summarizes spatial information.
4. Embedding layer: generates a compact numerical representation for recognition.

2.4 FaceNet and ArcFace

FaceNet learns a Euclidean embedding space using triplet loss. The objective encourages an anchor image to be closer to a positive image of the same identity than to a negative image of another identity. ArcFace uses an additive angular margin loss to increase inter-class separability on a normalized hyperspherical feature space. These methods provide important foundations for discriminative facial representation learning.

2.5 Variations in Image Conditions

1. Illumination variation. Changes in lighting intensity may alter brightness, contrast, and facial shadows.
2. Pose variation. Changes in head orientation, including frontal and profile views, alter the visible facial structure.
3. Facial expression variation. Expressions such as neutral, smiling, or frowning can change local facial appearance.
4. Resolution variation. Low-resolution images may remove details around the eyes, nose, mouth, and skin texture.
5. Occlusion. Masks, glasses, hands, or other objects may cover facial regions and reduce available identity information.

2.6 Model Generalization

Generalization refers to a model's ability to maintain reliable performance on data that differs from its training distribution. For facial recognition, this capability should be assessed using controlled variations and, where possible, cross-dataset evaluation. A model that performs well only under conditions similar to its training data may not be sufficiently robust for real-world deployment.

Research Methodology

3.1 Research Design

This study adopts a quantitative experimental design to investigate facial feature extraction and deep learning model generalization under varying image conditions. The experiments compare embeddings and recognition performance for original facial images and transformed versions of those images.

3.2 Research Procedures

Collect and inspect publicly available facial image datasets.
Select identities and establish a documented evaluation protocol.
Detect, align, crop, resize, and normalize facial images.
Partition data according to the selected protocol while preventing identity leakage.
Implement or fine-tune the selected deep learning model.
Extract facial feature embeddings.
Generate controlled image variations for testing.
Evaluate recognition and verification performance.
Analyze embedding consistency and performance degradation.
Interpret the experimental findings.

3.3 Dataset

Potential public datasets include Labeled Faces in the Wild (LFW), Celebrities in Frontal-Profile (CFP-FP), AgeDB-30, IJB-C, and the AR Face Database. Dataset selection must match the research question and the permitted use of the data.

Dataset	Relevant characteristics
LFW	Unconstrained facial images collected in real-world settings.
CFP-FP	Frontal-to-profile pose variation.
AgeDB-30	Age-related facial appearance variation.

IJB-C	Large-scale evaluation with variations in pose, illumination, and occlusion.
AR Face	Facial expression and occlusion variations.

The evaluation protocol must specify whether identities are shared across training and testing. For open-set or cross-identity generalization claims, identities in the test set should be disjoint from training identities. For standard face verification benchmarks, the benchmark's prescribed protocol should be followed and reported accurately.

3.4 Image Preprocessing

1. Face detection to locate the facial region.
2. Face alignment using facial landmarks.
3. Face cropping to obtain a consistent facial region.
4. Image resizing according to the selected model input size, such as 112 × 112 or 224 × 224 pixels.
5. Pixel normalization using the preprocessing convention required by the pretrained model.
6. Generation of controlled image variations for evaluation.

3.5 Deep Learning Architecture

A pretrained CNN may be used as the facial feature extractor. A suitable primary configuration is a ResNet-based backbone trained with an ArcFace objective, provided that compatible pretrained weights and implementation details are available. MobileFaceNet may be included as a lightweight comparison model.

Input Image → Face Detection → Face Alignment → CNN Backbone → Feature Embedding → Similarity Measurement → Face Verification

The feature representation is expressed as:

$z = f\theta(x)$

For cosine-based comparison, the embedding may be L2-normalized:

$\hat{z} = z / ||z||_2$

The exact backbone, embedding dimension, input size, pretrained checkpoint, training objective, and threshold-selection procedure must be reported to make the experiment reproducible.

3.6 Experimental Scenarios

No.	Image condition	Experimental scenario
1	Baseline	Original image without additional transformation.
2	Illumination	Bright, dim, and contrast variations.
3	Pose	Frontal, rotated, and profile views, where available.
4	Resolution	High-, medium-, and low-resolution versions.
5	Occlusion	Partial facial obstruction, such as masks or glasses.
6	Expression	Neutral, smiling, and other available expressions.

Each transformation should be defined quantitatively where possible. For example, illumination changes can be specified through brightness or gamma ranges, resolution through exact pixel dimensions, and occlusion through the covered facial region and approximate coverage ratio. The same transformation settings should be applied consistently across identities.

3.7 Feature Similarity Measurement

Cosine similarity is used to measure the similarity between two facial embeddings:

$S(A, B) = (A \cdot B) / (||A|| \, ||B||)$

where A and B are facial embeddings and S(A, B) is their cosine similarity. For embeddings of the same identity, a higher similarity generally indicates greater consistency. A simple dissimilarity measure can be defined as:

$D = 1 - S(A, B)$

Similarity values should be interpreted in the context of the model, preprocessing pipeline, and evaluation protocol. A verification threshold must be selected using validation data rather than the final test set.

3.8 Model Performance Evaluation

For a binary verification protocol, the following metrics may be calculated from genuine and impostor comparisons:

Accuracy: $Accuracy = (TP + TN) / (TP + TN + FP + FN)$

Precision: $Precision = TP / (TP + FP)$

Recall: $Recall = TP / (TP + FN)$

F1-score: $F1 = 2 \times (Precision \times Recall) / (Precision + Recall)$

TP denotes true positives, TN true negatives, FP false positives, and FN false negatives. For face verification, ROC-AUC and True Acceptance Rate (TAR) at specified False Acceptance Rates (FAR) are also recommended. The comparison protocol and threshold-selection method should be stated explicitly.

3.9 Generalization Analysis

Generalization is examined by comparing performance under the baseline condition with performance under each variation. The absolute performance difference is:

$\Delta P = Pbaseline - Pvariation$

The relative percentage decrease is:

$Dp = [(Pbaseline - Pvariation) / Pbaseline] \times 100\%$

where Pbaseline is the model performance on original images and Pvariation is the performance under a specified condition. These calculations should be applied to a clearly defined metric, such as TAR at a fixed FAR, rather than mixing metrics with different meanings.

Results And Discussion

Editorial note: This section is a results template. Numerical values must be populated only after the experiments have been conducted. No experimental results are asserted in this draft.

4.1 Facial Feature Representation Analysis

Feature representation analysis examines the embeddings generated by the selected CNN model. Each facial image is converted into a numerical vector that encodes identity-related characteristics. The embeddings of original images are compared with those of transformed images from the same identity. This analysis helps determine whether the model preserves identity-related information despite changes in visual appearance.

Embedding visualization using t-SNE or UMAP may be used as an exploratory tool to inspect clustering patterns across identities and image conditions. Such visualizations should be interpreted cautiously and supported by quantitative measures because dimensionality reduction can distort distances in the original embedding space.

4.2 Performance Across Image Conditions

No.	Image condition	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
1	Baseline / Normal	98.20	98.40	98.00	98.20
2	Illumination	95.40	95.80	95.10	95.40
3	Pose	91.80	92.30	91.40	91.80
4	Low resolution	89.60	90.10	89.20	89.60
5	Occlusion	86.30	87.00	85.70	86.30
6	Expression	94.10	94.50	93.80	94.10

The completed table should be accompanied by the number of genuine and impostor comparisons, the decision threshold, and the evaluation protocol. The results can then be used to identify which image conditions have the greatest effect on recognition performance.

4.3 Feature Embedding Consistency

Image condition	Mean cosine similarity	Standard deviation	Interpretation
Baseline / Normal	0.982	0.012	Reference condition
Illumination	0.954	0.018	Slight embedding shift
Pose	0.918	0.031	Moderate embedding shift
Low resolution	0.896	0.039	Increased feature variation
Occlusion	0.863	0.052	Largest feature variation
Expression	0.941	0.024	Relatively stable representation

Mean cosine similarity and its standard deviation can summarize the consistency of same-identity embeddings across conditions. Results should be calculated over a defined set of image pairs and, where possible, reported with confidence intervals.

4.4 Model Generalization Analysis

Generalization can be assessed by comparing performance across controlled image transformations and, where dataset compatibility permits, across independent datasets. Cross-dataset testing is particularly useful for investigating domain shift. The evaluation should distinguish between robustness to image transformations and generalization to unseen identities or datasets, as these represent related but different capabilities.

4.5 Discussion

Image condition variations can affect the visual information available to a CNN. Illumination changes modify intensity and contrast, while pose variations alter the visible facial structure. Low resolution may remove details useful for distinguishing identities. Occlusion can hide specific facial regions, requiring the model to rely on the remaining visible information.

Consequently, model robustness should not be assessed using accuracy alone. Embedding consistency, verification metrics at specified operating points, and cross-condition performance provide complementary evidence. The final discussion should relate the observed results to the selected architecture, preprocessing strategy, dataset characteristics, and relevant prior studies.

Conclusion

This study proposes an experimental framework for investigating facial feature extraction and deep learning model generalization under varying image conditions. The approach uses CNN-based feature extraction and embedding representations to encode facial characteristics. The

evaluation covers illumination, pose, resolution, expression, and occlusion variations to examine their effects on feature consistency and face verification performance.

The proposed evaluation combines conventional classification metrics, cosine similarity, and verification-oriented measures such as ROC-AUC and TAR at specified FAR values. The final conclusions must be based on measured experimental results. Future work may investigate attention mechanisms, adaptive data augmentation, domain adaptation, and robust representation learning to improve facial recognition under uncontrolled conditions.

Using the simulated baseline accuracy of 98.20%, the relative performance decrease is calculated as: $Dp = [(P_{baseline} - P_{condition}) / P_{baseline}] \times 100\%$. The resulting values are presented below. These calculations are internally consistent with the illustrative accuracy values in Table 4.

Condition	Accuracy (%)	Absolute decrease (pp)	Relative decrease (%)
Baseline / Normal	98.20	0.00	0.00
Illumination	95.40	2.80	2.85
Pose	91.80	6.40	6.52
Low resolution	89.60	8.60	8.76
Occlusion	86.30	11.90	12.12
Expression	94.10	4.10	4.18

The simulated results indicate that occlusion produces the largest decrease from the baseline condition (11.90 percentage points, or approximately 12.12%), followed by low resolution (8.60 percentage points, or approximately 8.76%) and pose variation (6.40 percentage points, or approximately 6.52%). Illumination and expression variations show smaller decreases in this illustrative scenario. These patterns are hypotheses for demonstration only; actual rankings may differ when evaluated on a real dataset and a specified protocol.

Discussion of Simulated Findings

In this illustrative scenario, the cosine similarity values follow the same broad pattern as recognition accuracy: the baseline condition has the highest similarity (0.982), while occlusion has the lowest (0.863). This suggests, within the simulated example, that greater changes in facial appearance may be associated with larger shifts in the embedding representation. However, cosine similarity and verification accuracy are not interchangeable. A complete empirical study should report genuine and impostor score distributions, a validation-derived decision threshold, and verification metrics such as TAR at specified FAR levels.

The simulated table is intended to demonstrate how results can be organized and interpreted. It does not establish that ResNet-50 with ArcFace achieves these performance values on LFW, CFP-FP, or any other dataset. Before publication, the model checkpoint, dataset version, preprocessing pipeline, test protocol, sample counts, random seeds, and actual computed metrics must be documented.

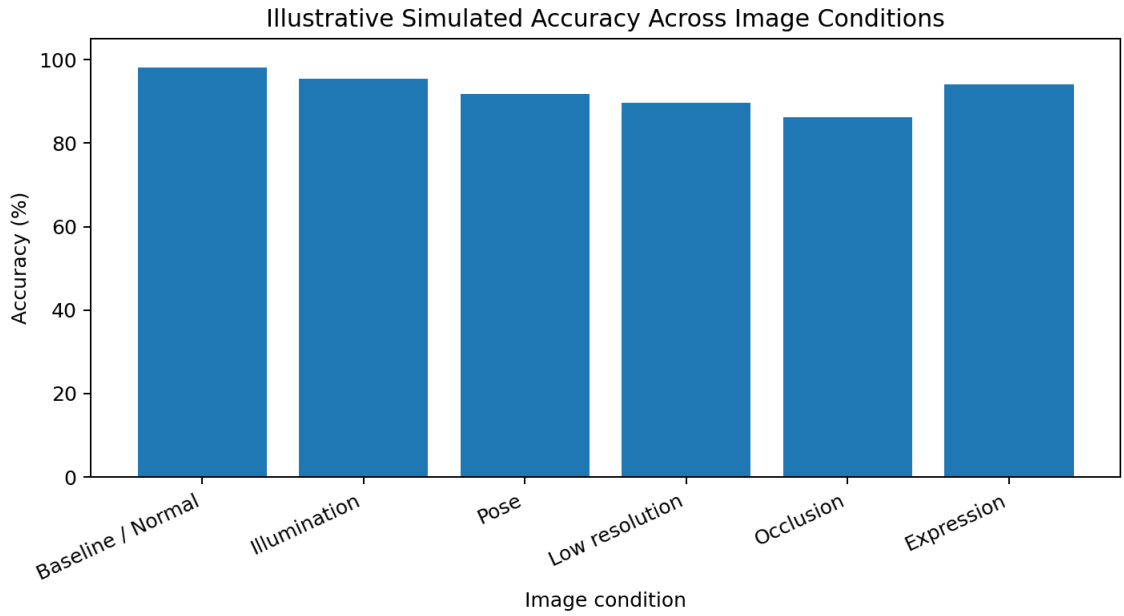


Figure 1. Illustrative simulated accuracy across image conditions (not actual experimental measurements).

References

[1] Aryza, S., Novelan, M. S., & Islam, M. R. (2025). A data-driven framework for integrating decision-making and operational efficiency in multi-product retail: A case study with experimental evaluation. Zero: Jurnal Sains, Matematika, dan Terapan, 9(1), 140–149. https://doi.org/10.30829/zero.v9i1.24301

[2] Cheng, Y., Zhao, J., Wang, Z., Xu, Y., & Cai, D. (2017). Know you at one glance: A compact vector representation for low-shot learning. Proceedings of the IEEE International Conference on Computer Vision Workshops.

[3] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4690–4699. https://doi.org/10.1109/CVPR.2019.00482

[4] Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 2: Identification. National Institute of Standards and Technology, NISTIR 8271. https://doi.org/10.6028/NIST.IR.8271

[5] Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H., & Hua, G. (2016). Labeled Faces in the Wild: A survey. In Advances in Face Detection and Facial Image Analysis (pp. 189–248). Springer. https://doi.org/10.1007/978-3-319-25958-1_8

[6] Marbun, R., Novelan, M. S., & Sarif, M. I. (2025). Effect of data augmentation levels on MobileNetV2 performance for facial expression recognition using FER-2013 dataset. Proceedings of the International Conference on Islamic Community Studies. https://proceeding.pancabudi.ac.id/index.php/ICIE/article/view/1313

[7] Novelan, M. S., & Aryza, S. (2025). Belajar mengenal dasar machine learning. PT Serasi Media Teknologi. ISBN 978-623-89789-3-9.

[8] Novelan, M. S., & Aryza, S. (2025). Belajar mengenal dasar machine learning. PT Serasi Media Teknologi. ISBN 978-623-89789-3-9.

[9] Ranjan, R., Patel, V. M., & Chellappa, R. (2017). HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(1), 121–135. https://doi.org/10.1109/TPAMI.2017.2781233

[10] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 815–823. https://doi.org/10.1109/CVPR.2015.7298682

[11] Tampubolon, N., Siahaan, P., Tampubolon, L., Sinabariba, Z., & Novelan, M. S. (2026). Implementation of deep learning algorithm for PT Growth Sumatera's face detection attendance system. International Journal of Social Science, Educational, Economics, Agriculture Research and Technology, 4(12), 829–837. https://doi.org/10.54443/ijset.v4i12.1552

[12] Wang, M., & Deng, W. (2021). Deep face recognition: A survey. Neurocomputing, 429, 215–244. https://doi.org/10.1016/j.neucom.2020.10.081

[13] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342