

Influence of Minkowski Distance Order on Deep Embedding-Based Cross-Spectral Cross-Distance Face Recognition

Ernita Dewi Meutia^{1,2}, Rusdha Muharar², Nasaruddin Nasaruddin², Fitri Arnia²

¹ Doctoral Program, School of Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia

² Department of Electrical and Computer Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia

Corresponding author: Fitri Arnia (e-mail: f.arnia@usk.ac.id), **Author(s) Email:** Ernita Dewi Meutia (e-mail: ernita.dmeutia@usk.ac.id), Nasaruddin (e-mail: nasaruddin@usk.ac.id), Rusdha Muharar (email: r.muharar@usk.ac.id)

Abstract: Cross-spectral cross-distance face recognition (CSCD-FR) is a challenging biometric task involving visible (VIS) and near-infrared (NIR) facial images acquired at varying stand-off distances. Beyond the spectral discrepancy between VIS and NIR images, increasing acquisition distance degrades resolution, blurs images, and reduces facial detail, making recognition considerably more difficult. Although deep convolutional neural networks (CNNs) have demonstrated strong capabilities for learning discriminative facial embeddings, existing studies have predominantly focused on network architectures, feature representations, and learning strategies, while the influence of distance metric selection remains relatively underexplored. Since recognition decisions are ultimately based on distances between embedding vectors, the choice of similarity metric may affect recognition performance. This study therefore investigates the influence of generalized Minkowski distance order on CSCD-FR using deep embeddings extracted from MobileNetV3-Small, EfficientNet-B2, and DenseNet121. Using the Long-Distance Heterogeneous Face Database (LDHF-DB) at stand-off distances of 1, 60, and 100 m, both ImageNet-pretrained and domain-adapted models were evaluated using L_2 -normalized embeddings. Domain adaptation optimized the embedding layer using triplet margin loss while the CNN backbones remained frozen. Minkowski orders from $p = 1$ to $p = 7$ were evaluated using closed-set identification and verification protocols. ImageNet-pretrained models yielded Rank-1 accuracies below 11%, whereas domain adaptation substantially improved recognition performance. DenseNet121 achieved the highest observed performance, with 93.19% Rank-1 accuracy and 2.16% EER at $p = 5$. The Rank-1-optimal Minkowski distance order varied across architectures, with EfficientNet-B2 and MobileNetV3-Small achieving their highest observed Rank-1 accuracies at $p = 7$ and $p = 5$, respectively. These findings show that selecting a similarity metric can provide an additional matching-stage optimization without modifying the CNN backbone. Because the same retained identities were used for domain adaptation and evaluation, the results represent a dataset-specific assessment rather than recognition performance on previously unseen subjects.

Keywords: Face recognition; cross-spectral; cross-distance; Minkowski distance norm; deep embedding.

1. Introduction

Face recognition is a widely used biometric technology for non-intrusive identity verification in applications such as surveillance, access control, forensic investigation, bank authorization, and public safety [1][2]. Deep convolutional neural networks (CNNs) have significantly improved recognition accuracy by learning robust facial features from large datasets [3]. However, reliable face recognition under real-world conditions remains challenging. Variations in pose, occlusion, lighting, imaging modality, and acquisition distance can all alter facial appearance and make recognition more

difficult [4][5]. To enable face recognition in low-light and nighttime conditions, near-infrared (NIR) imaging is widely adopted because it is less sensitive to ambient illumination than conventional visible (VIS) imaging. Recognition systems that match NIR probe images against VIS gallery images are commonly referred to as cross-spectral face recognition (CSFR) [6][7][8]. In practical surveillance and security applications, however, recognition systems must also operate across different stand-off distances rather than under a fixed imaging condition. As acquisition distance increases, facial images can lose spatial detail because of reduced resolution, optical blur, and sensor

limitations [9]. When these distance-related degradations are combined with the inherent appearance differences between VIS and NIR images, the recognition problem becomes considerably more challenging. This combined problem is referred to as cross-spectral cross-distance face recognition (CSCD-FR) [10] [11].

Research on this challenging setting has explored different strategies to address spectral and acquisition-distance variations. [10][12] developed the Long-Distance Heterogeneous Face Database (LDHF-DB), which provides images captured under different spectral and acquisition-distance conditions for investigating heterogeneous face recognition. [13] employed homomorphic filtering for photometric normalization and Band-Limited Phase-Only Correlation (BLPOC) for phase-based image matching. More recent NIR-VIS face recognition studies have increasingly focused on deep feature representations and domain adaptation. [14] proposed a semi-supervised pseudo-label association and prototype-based invariant learning to improve domain-invariant identity representations, while [15] demonstrated the effectiveness of large VIS-pretrained networks and regularized fine-tuning for NIR-VIS face recognition. These approaches primarily address spectral and domain variations through improved feature learning, representation transfer, or domain adaptation. More recently, in CSCD-FR, [16] investigated NIR-to-VIS image translation using an asymmetric CycleGAN and transfer learning, followed by a framework combining image enhancement, modified CycleGAN-based translation, and transfer learning on a newly introduced CSCD-F dataset [17]. These studies primarily focus on improving image transformation, domain adaptation, or learned feature representations.

Despite significant advances in feature extraction networks and loss functions, the similarity measurement stage has received comparatively less attention. Modern face recognition systems compare feature embeddings rather than raw facial images; therefore, recognition performance depends not only on the discriminative quality of the embeddings but also on the distance metric used to compare them. Even highly discriminative embeddings may yield suboptimal performance when paired with an unsuitable similarity measure [18]. Euclidean distance and cosine similarity are widely used because they are computationally simple, well-established, and well-suited for comparing normalized deep embeddings [19] [20]. Euclidean distance is naturally related to the geometric structure of the embedding space, while cosine similarity measures the angular relationship between normalized embeddings, making both measures particularly convenient for deep feature comparison. Consequently, they have become dominant choices in deep face recognition and related metric-learning

systems [21] [22]. However, different distance metrics can produce different recognition performance even when applied to the same feature embeddings [18]. In CSCD-FR, previous studies have similarly adopted specific matching mechanisms, such as BLPOC [13] or Euclidean feature distance [12] [17], but have generally treated the matching measure as a fixed component rather than an independent experimental variable. Thus, the effect of the matching distance itself remains insufficiently explored in CSCD-FR.

The generalized Minkowski distance provides a systematic framework for investigating this issue. It represents a family of distance measures in which Manhattan and Euclidean distance correspond to specific distance orders [23], allowing the effect of the distance order to be evaluated while keeping the underlying feature representation unchanged. Although distance selection has long been recognized as an important consideration in pattern recognition [18] [23] [24], the influence of different Minkowski orders on learned deep facial embeddings has not been systematically investigated under simultaneous cross-spectral and cross-distance conditions. In particular, it remains unclear whether the conventional Euclidean distance is optimal for deep embeddings in CSCD-FR.

The aim of this study is to systematically investigate the influence of the generalized Minkowski distance on deep embedding-based CSCD-FR, analysing how different distance orders affect identification and verification performance. Rather than proposing a new network architecture, image-enhancement method, or image-translation framework, the study focuses specifically on the matching stage. Using the LDHF-DB dataset, facial embeddings are extracted using MobileNetV3-Small [25], EfficientNet-B2 [26], and DenseNet121 [27]. Their matching performance is evaluated using different Minkowski distance orders. Both ImageNet-pretrained and domain-adapted models are examined to compare generic pretrained representations with embeddings adapted to the CSCD-FR domain. The main contributions of this paper are twofold:

1. We systematically investigate the generalized Minkowski distance on deep embedding-based CSCD-FR, analysing how different distance orders affect identification and verification performance using domain-adapted CNN embeddings. ImageNet-pretrained embeddings are also included as baselines to examine how domain adaptation affects the behavior of similarity metrics.
2. We perform a controlled comparison of DenseNet121, MobileNetV3-Small, and EfficientNet-B2 using the same preprocessing and evaluation methods. The results show that the Rank-1-optimal Minkowski order varies across CNN

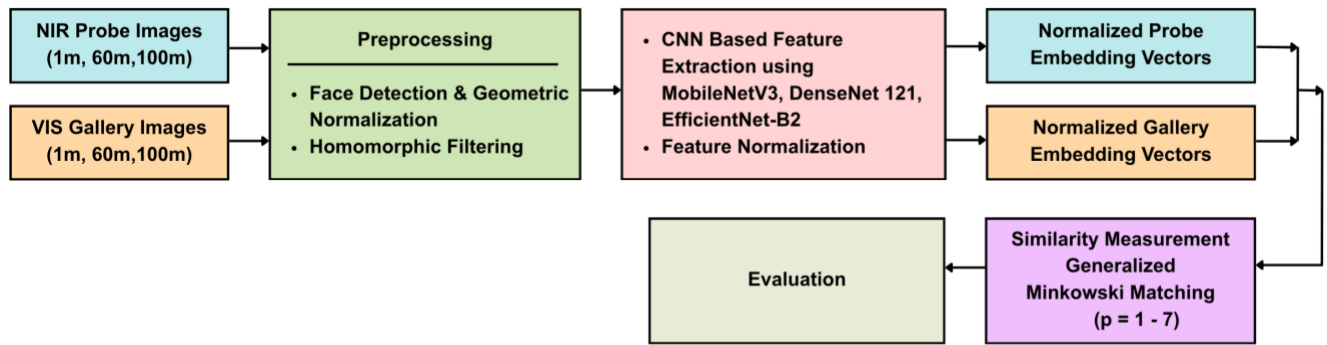


Fig. 1. Overview of the proposed cross-spectral cross-distance face recognition framework.

Table 1. Summary of CNN architectures and model configurations.

Model	Parameters	Embedding	Inference Time (ms/image)	FLOPS
MobileNetV3-Small	~1.22 M	512-D	283	112 M
EfficientNet-B2	~8.42 M	512-D	405	1.158 G
DenseNet121	~7.48 M	512-D	450	2.897 G
ResNet50	~24.56 M	512-D	318	8.2 G

architectures, highlighting the interaction between learned feature representations and distance metrics in CSCD-FR. The remainder of the paper is organized as follows: Section II explains the method, including the database setup, domain adaptation, and the evaluation scenario. Sections III and IV present the empirical results and discussion. Section V provides the conclusion.

II. Method

The experimental workflow is illustrated in Fig. 1. The proposed CSCD-FR framework comprises four stages: preprocessing, feature extraction, feature matching, and performance evaluation. VIS gallery and NIR probe images first undergo face detection, geometric normalization, homomorphic filtering, and resizing. The preprocessed images are then processed by CNNs to extract 512-dimensional facial embeddings, followed by L_2 -normalization, to project them onto a common hypersphere. The normalized embeddings are subsequently compared using the generalized Minkowski distance, and recognition performance is evaluated using closed-set identification and verification protocols.

The primary experimental variable is the generalized Minkowski distance order, which is systematically varied from $p = 1$ to $p = 7$ while keeping preprocessing, feature extraction, embedding normalization, and evaluation procedures unchanged. This controlled design allows differences in recognition performance to be attributed to the matching distance rather than to changes in the preceding stages of the pipeline. The complete experimental procedure is

summarized in Algorithm 1, and the individual methodological stages are described in the following sections.

Algorithm 1. Summary of the experimental procedure.

Input: Preprocessed VIS gallery and NIR probe images; CNN models; Minkowski orders ($p = 1$ to 7).

1. Preprocess VIS and NIR images through face detection, geometric normalization, grayscale conversion, homomorphic filtering, and resizing to (224×224) .
2. Extract 512-dimensional embeddings from the VIS and NIR images using the selected CNN, and apply (L_2) normalization to all embeddings.
3. For each Minkowski order ($p = 1$ to 7):
 - a. Compute distances between every NIR probe and VIS gallery embedding.
 - b. Rank gallery images by increasing distance and calculate Rank-1 accuracy and CMC.
 - c. Construct genuine and impostor comparisons and calculate ROC, AUC, and EER.
4. For each CNN, identify the evaluated Minkowski order producing the highest observed Rank-1 accuracy.
5. Compare the selected order with ($p = 2$) using McNemar's test for Rank-1 outcomes and bootstrap analysis for EER, with 95% confidence intervals.
6. Report the identification, verification, and statistical results.

A. Dataset

This study uses the publicly available Long-Distance Heterogeneous Face Database (LDHF-DB) [12], which contains VIS and NIR facial images acquired at 1 m, 60

m, 100 m, and 150 m. The 150 m subset was excluded because the Viola–Jones face detector used in the preprocessing pipeline failed to detect faces in most of images acquired at this distance, preventing reliable face extraction for subsequent feature representation and matching. This limitation is consistent with the observations reported by Kang et al. [10], who noted that facial images acquired at 150 m lose most facial details, making both face detection and recognition considerably more challenging. Consequently, the present study was limited to the 1 m, 60 m, and 100 m subsets, for which reliable face detection and preprocessing could be consistently achieved.

Face detection on the retained subsets resulted in the exclusion of seven identities because one or more required face images could not be detected: Identity 0062 (60 m VIS and 60 m NIR), Identity 0092 (60 m NIR), and Identities 0014, 0041, 0081, 0082, and 0090 (100 m NIR). To maintain a complete probe–gallery protocol, any identity with missing images was removed entirely. The final dataset comprised 93 identities (558 images), each contributing one VIS gallery image and one NIR probe image at each acquisition distance (1 m, 60 m, and 100 m), yielding 93 genuine pairs and 8,556 impostor pairs for each probe–gallery distance combination. Accordingly, the exclusion of the 150 m subset represents a limitation of this study, as the reported results do not reflect performance under the most challenging long-range acquisition condition.

B. Data preprocessing

The preprocessing stages follow the CSCD-FR preprocessing pipeline described in [7]. All images underwent face detection to ensure that only facial regions were processed in subsequent stages. The Viola–Jones detector [28] was employed because it is computationally efficient, requires no additional training, and provides reliable performance for the retained 1 m, 60 m, and 100 m image subsets. As with other classical face detectors, its performance degrades for extremely small or low-quality faces, which contributed to the exclusion of the 150 m subset.

The detected faces were geometrically normalized using estimated eye locations for alignment, cropped, converted to grayscale to provide a common representation for both modalities, and resized to 224×224 pixels. To reduce illumination differences between VIS and NIR images while enhancing facial details, Homomorphic filtering was applied to all cropped facial images before feature extraction. The filter transfer function is given by Eq. (1) [13]:

$$H(u, v) = (\gamma_h - \gamma_l) \left[1 - \exp \left(-c \frac{D(u, v)^2}{D_0^2} \right) \right] + \gamma_l \quad (1)$$

where $D(u, v)$ is the distance from the frequency-domain origin, γ_l and γ_h control the amplification of low and high-frequency components, respectively, c

determines the sharpness of the transition, and D_0 represents the cutoff frequency.

The same Homomorphic filtering procedure was applied to both VIS and NIR images. To account for increased image degradation at longer acquisition distances, the low-frequency gain γ_l was adjusted for the longer-distance images, while the remaining parameters were kept unchanged. For images captured at 1 m, the filtering parameters were $\gamma_l = 0.85$, $\gamma_h = 2.0$, $c = 1.0$, and $D_0 = 30$. For images captured at 60 m and 100 m, $\gamma_l = 0.50$ was used, while the other parameters remained unchanged. These parameters were empirically determined during dataset preparation and held constant throughout all experiments. Consequently, the same pre-processed images were used across all CNN architectures and Minkowski distance orders, allowing performance differences to be evaluated under consistent preprocessing conditions. Representative examples of the filtered images are shown in Fig. 2. This preprocessing strategy requires prior knowledge of the acquisition distance and may not apply directly in operational environments where distance information is unavailable. Developing distance-independent preprocessing methods or automatic parameter selection remains an important direction for future research.

C. Feature Extraction and Normalization

Four CNN architectures were initially considered: ResNet50 [29], DenseNet121, EfficientNet-B2, and MobileNetV3-Small, representing different CNN design

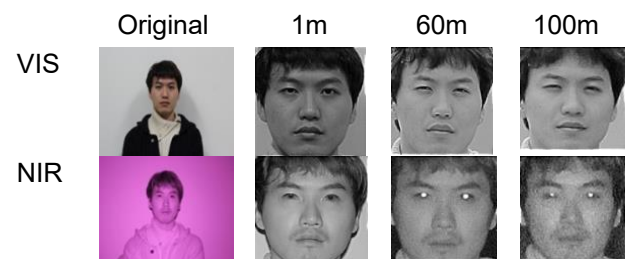


Fig. 2. Representative examples of filtered VIS and NIR images captured at 1 m, 60 m, and 100 m.

paradigms and computational complexities. These architectural configurations are summarized in Table 1.

All models were initialized with ImageNet-pretrained weights. For the pretrained baseline, the original ImageNet classification layer was replaced with an identity mapping, and the resulting backbone representations were extracted without additional training, providing baseline performance without dataset-specific adaptation. MobileNetV3-Small, EfficientNet-B2, and DenseNet121 were further adapted to the target domain through embedding optimization, while the adapted ResNet50 results were excluded from the comparative analysis. Thus,

architecture-level conclusions are restricted to the three domain-adapted models.

For domain adaptation, the original ImageNet classification layer was replaced with a trainable 512-dimensional embedding layer, while the convolutional backbone remained frozen. The classification layer was removed because its outputs correspond to the 1,000 ImageNet object categories and are not directly suitable for face recognition. The embedding layer was optimized using Triplet Margin Loss on images from all 93 retained identities. Triplets were formed by randomly selecting an anchor image, a positive image from the same identity, and a negative image from a different identity, regardless of modality or acquisition distance. The loss is defined as Eq. (2) [22]:

$$L_{\text{triplet}} = \max(\|f_a - f_p\|_2 - \|f_a - f_n\|_2 + m, 0) \quad (2)$$

where f_a , f_p and f_n denote the anchor, positive, and negative embeddings, respectively, and m is the triplet margin. Training was performed using the Adam optimizer with a learning rate of 0.001, a batch size of 16, and 300 epochs. The triplet margin was set to 1.0. No data augmentation or early stopping was employed. Because the objective of the domain adaptation was to obtain dataset-specific embeddings for evaluating similarity metrics rather than to assess generalization to unseen identities, no subject-disjoint training, validation, or testing partition was used.

All extracted probe and gallery embeddings were L_2 -normalized according to Eq.(3) and Eq.(4) [30] to produce unit-length representations.

$$\hat{x} = \frac{x}{\|x\|_2} \quad (3)$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2} \quad (4)$$

where x denotes the original embedding vector, $\hat{x}$ is the normalized embedding, x_i denotes the i -th element of the embedding vector, and d denotes the embedding dimensionality. L_2 -normalization projects embeddings onto a unit hypersphere, providing a consistent feature scale and reducing the influence of embedding magnitude during similarity computation [30]. The normalized embeddings were subsequently used for Minkowski distance matching. All experiments were implemented in PyTorch 2.11.0+cu128 and executed on Google-Colab using an NVIDIA Tesla T4 GPU.

D. Matching and Recognition

Similarity matching was performed using the generalized Minkowski distance. For each pair of a probe p_i and gallery embedding g_j , the distance was computed using Eq. (5) [23]:

$$D_p(p_i, g_j) = \left(\sum_{k=1}^d |p_{i,k} - g_{j,k}|^p \right)^{1/p} \quad (5)$$

where d is the embedding dimensionality, and p is the Minkowski distance order. Two commonly used special cases are the Manhattan distance ($p = 1$) and the Euclidean distance ($p = 2$) [23].

Identification was conducted under a closed-set protocol by comparing each NIR probe embedding with all VIS gallery embeddings. The predicted identity was assigned to the gallery sample having the minimum distance from the corresponding NIR probe embedding as defined in Eq. (6):

$$\hat{y}_i = \arg \min_{g_j \in \mathcal{G}} D_p(\hat{p}_i, \hat{g}_j) \quad (6)$$

where ($\mathcal{G}$) denotes the set of gallery embeddings, $\hat{p}_i$ denotes the normalized embedding of the i -th NIR probe image, and $\hat{g}_j$ denotes the normalized embedding of the j -th gallery image. The gallery sample yielding the smallest distance was selected as the predicted identity [31].

To evaluate the impact of the similarity metric, the Minkowski distance order was systematically varied across ($p \in 1, 2, 3, 4, 5, 6, 7$). The range includes the Manhattan and Euclidean distances as well as higher-order norms. The range was selected to examine the effect of increasing distance order, without extending the analysis to unnecessarily large values that showed no additional benefit in preliminary experiments. Preprocessing, feature extraction, and evaluation procedures were kept constant across all orders, so that observed performance differences could be primarily attributed to the similarity metric.

E. Recognition Performance Evaluation

The face recognition system was evaluated using two complementary protocols: closed-set identification (1:N), which served as the primary evaluation, and verification (1:1) [32]. Identification was measured using overall Rank-1 accuracy, distance-specific performance across all nine probe-gallery distance combinations formed by the 1 m, 60 m, and 100 m acquisition distances, and cumulative match characteristic (CMC) curves (Rank-1 to Rank-20). Verification was assessed as a supplementary evaluation using the equal error rate EER [33]. Overall identification performance was evaluated using 279 NIR probe images and 279 VIS gallery images, resulting in 279 identification searches and 77,841 probe-gallery comparisons. Rank-1 accuracy and CMC curves were computed from these searches. To analyze the effect of acquisition distance, the evaluation was repeated separately for each of the nine probe-gallery distance combinations formed by the 1 m, 60 m, and 100 m subsets. Each combination consisted of 93 probe images and 93 gallery images, yielding 93 identification searches and 8,649 probe-gallery comparisons. The resulting Rank-1 accuracies were summarized in a 3×3 performance matrix.

The rank- k recognition rate was computed using Eq.(7) [34]:

$$\text{Rank-}k = \frac{1}{N} \sum_{i=1}^N 1(r_i \leq k) \quad (7)$$

where r_i is the rank of the correct gallery identity for the i -th probe, N denotes the total number of probe images,

and $\mathbb{1}(\cdot)$ is the indicator function, which returns 1 when the correct identity appears within the top k retrieved gallery candidates ($r_i \leq k$) and 0 otherwise. Rank-1 accuracy is obtained when $k=1$, indicating that the correct gallery identity is the top-ranked gallery match.

The distance-specific evaluation was divided into same-distance analysis (SDA) and cross-distance analysis (CDA). SDA refers to probe and gallery images acquired at the same distance, whereas CDA refers to probe and gallery images acquired at different distances. For each probe-gallery distance combination, the Rank-1 identification accuracy was computed using Eq. (7) with $k = 1$. Overall SDA and CDA were then calculated by averaging the corresponding distance-specific Rank-1 accuracies, as defined in Eq. (8) and Eq. (9):

$$SDA = \frac{1}{N_d} \sum_{i=1}^{N_d} A(d_i, d_i) \quad (8)$$

$$CDA = \frac{1}{N_d(N_d - 1)} \sum_{i=1}^{N_d} \sum_{j=1, j \neq i}^{N_d} A(d_i, d_j) \quad (9)$$

where $A(d_i, d_j)$ denotes the Rank-1 accuracy when probe images acquired at distance d_i are matched against gallery images acquired at distance d_j , and N_d is the number of acquisition distances. SDA and CDA are study-defined aggregation measures introduced to summarize same-distance and cross-distance identification performance, respectively.

Verification performance was evaluated using the complete set of probe-gallery comparisons, yielding 837 genuine pairs and 77,004 impostor pairs for each CNN model and Minkowski distance order. These scores were pooled to construct receiver operating characteristic (ROC) curves. Because the Minkowski distance is a dissimilarity measure, lower distances indicate a higher likelihood of a genuine match. The decision threshold was swept across the full range of observed distances to compute the false acceptance rate (FAR) and false rejection rate (FRR), as defined in Eq. (10) and Eq. (11) [31]. Because the system uses distance-based matching, a trial is accepted when its distance is less than or equal to the decision threshold. Accordingly, an impostor trial with a distance below or equal to the threshold constitutes a false acceptance, whereas a genuine trial with a distance greater than the threshold constitutes a false rejection.

$$FAR = \frac{\text{Number of impostor score} \leq \text{threshold}}{\text{Total number of impostor trials}} \quad (10)$$

$$FRR = \frac{\text{Number of genuine score} > \text{threshold}}{\text{Total number of genuine trials}} \quad (11)$$

The equal error rate (EER) was estimated at the operating point where FAR and FRR were equal, or as close as possible for the discrete score distribution. The EER was determined according to Eq.(12) [13]:

$$EER = FAR(t^*) \approx FRR(t^*) \quad (12)$$

where t^* denotes the threshold at which the absolute difference between FAR and FRR is minimized. In addition to EER, the area under the receiver operating characteristic curve (AUC) and the true acceptance rate (TAR) at FAR values of 1% and 0.1% were reported. TAR is defined as Eq.(13) [31]:

$$TAR(t) = 1 - FRR(t) = \frac{N_{genuine}(D \leq t)}{N_{genuine}} \quad (13)$$

where t is the decision threshold. $N_{genuine}(D \leq t)$ denotes the number of genuine trials with distances less than or equal to the threshold, and $N_{genuine}$ denotes the total number of genuine trials. The AUC was calculated as the area under the ROC curve, defined as Eq.(14) [34]:

$$AUC = \int_0^1 TAR(FAR) d(FAR) \quad (14)$$

TAR values were specifically reported at FAR = 1% and 0.1% to characterize verification performance at increasingly stringent false-acceptance operating points.

F. Statistical Analysis

Statistical analyses were performed to determine whether the performance differences between the Euclidean distance ($p = 2$), and the Rank-1-optimal Minkowski distance order for each domain-adapted CNN model were statistically significant. For identification, 95% confidence intervals (CIs) for Rank-1 accuracy were estimated using nonparametric bootstrap resampling with replacement. Because the Euclidean distance and the Rank-1-optimal Minkowski distance order were evaluated on the same set of probe images, their identification outcomes were paired. Pairwise differences in Rank-1 performance were therefore evaluated using the exact McNemar test. Under the null hypothesis Eq.(15), the probabilities of the two discordant outcomes are equal [35]:

$$H_0: p_{01} = p_{10} \quad (15)$$

where p_{01} and p_{10} denote the probabilities of the two discordant outcomes. Conditional on the total number of discordant observations, the number of observations in one discordant category follows a binomial distribution under the null hypothesis, as given in of Eq.(16) [35]:

$$X \sim \text{Binomial}(n_{01} + n_{10}, 0.5) \quad (16)$$

where n_{01} and n_{10} are the observed frequencies of the two discordant outcomes. The two-sided exact p -value was obtained from this binomial distribution. Statistical significance was assessed at ($\alpha = 0.05$). For verification, a nonparametric paired bootstrap with 5,000 replicates was performed at the probe level following the bootstrap framework in [36]. For each model comparison, the 279 NIR probes were sampled with replacement, and the same resampled probes were evaluated using the Euclidean distance and the

Rank-1-optimal Minkowski distance order. This paired procedure preserved the correspondence between the two distance metrics across bootstrap samples. For each bootstrap replicate b , the differences in Rank-1 accuracy and EER were calculated using Eq.(17) and Eq.(18), respectively:

$$\Delta R^{(b)} = R_{p^*}^{(b)} - R_{p=2}^{(b)} \quad (17)$$

$$\Delta EER^{(b)} = EER_{p^*}^{(b)} - EER_{p=2}^{(b)} \quad (18)$$

The selected order p^* represents the highest observed Rank-1 performance among the evaluated orders under the adopted dataset-specific protocol, and b denotes the bootstrap replicate. Bootstrap resampling was then used to estimate the sampling distribution and 95% confidence intervals of the paired performance differences [36], as defined in Eq. (19):

$$CI_{95\%} = [Q_{0.025}(\Delta^*), Q_{0.975}(\Delta^*)] \quad (19)$$

where $Q_{0.025}$ and $Q_{0.975}$ represent the 2.5th and 97.5th percentiles of the corresponding bootstrap difference distribution Δ^* . The two-sided bootstrap p -value was estimated from the proportions of bootstrap differences falling on either side of the null value of zero as in Eq. (20) [36]:

$$p_{\text{boot}} = 2 \min \left[\frac{1}{B} \sum_{b=1}^B 1(\Delta^{(b)} \leq 0), \frac{1}{B} \sum_{b=1}^B 1(\Delta^{(b)} \geq 0) \right], \quad B = 5000. \quad (20)$$

A negative ΔEER indicates a reduction in EER relative to the Euclidean baseline, whereas a positive ΔR indicates an improvement in Rank-1 accuracy. The bootstrap procedure was applied using $p^* = 5$ for DenseNet121 and MobileNetV3-Small and $p^* = 7$ for EfficientNet-B2. Differences were considered statistically significant when $p < 0.05$.

III. Result

A. ImageNet-Pretrained Reference Performance

To establish a reference for the CSCD-FR task, ResNet50, DenseNet121, EfficientNet-B2, and MobileNetV3-Small were evaluated using ImageNet-pretrained weights without domain adaptation. Their embeddings were compared using Minkowski distance orders $p = 1$ to $p = 7$, with the Rank-1 results summarized in the pre-trained rows (PT) of Table 2.

Table 2. Overall Rank-1 Accuracy of pretrained (PT) and domain-adapted (DA) CNN models at different Minkowski distance orders (%).

Model	Type	p = 1	p = 2	p = 3	p = 4	p = 5	p=6	p=7
ResNet50	PT	8.96	8.60	7.88	6.09	5.73	5.02	5.02
DenseNet121	PT	9.68	9.32	7.89	8.60	8.24	7.52	7.52
EfficientNet-B2	PT	10.04	10.39	10.04	7.89	6.45	6.09	6.45
MobileNetV3-Small	PT	8.24	7.89	8.60	7.53	5.73	6.09	5.38
DenseNet121	DA	91.76	92.47	92.83	92.83	93.19	92.47	92.11
EfficientNet-B2	DA	79.57	80.29	80.64	82.08	82.44	83.87	84.23
MobileNetV3-small	DA	88.17	88.17	89.61	89.96	91.40	91.04	90.32

Among the pretrained models, EfficientNet-B2 achieved the highest Rank-1 accuracy of 10.39% at $p = 2$, followed by DenseNet121 at 9.68% at $p = 1$, ResNet50 at 8.96% at $p = 1$, and MobileNetV3-Small at 8.60% at $p = 3$. Across all four architectures and distance orders, Rank-1 accuracy remained below 11%, indicating limited identification capability when

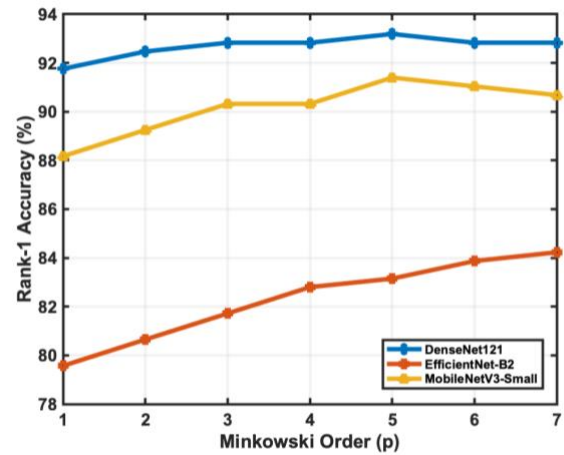


Fig. 3. Influence of Minkowski distance order on Rank-1 identification performance of domain-adapted CNN models.

ImageNet-pretrained embeddings were used without domain adaptation. The corresponding verification results in Table 3, show a similar limitation. EfficientNet-B2 achieved the lowest EER of 41.71% at $p = 1$, followed by DenseNet121 at 45.62%, ResNet50 at 45.74%, and MobileNetV3-Small at 46.48%, all at $p = 2$. Thus, all pretrained models exhibited high EER above 41%, confirming that generic ImageNet representations provided limited discriminative capability under the evaluated CSCD-FR conditions.

B. Domain-Adapted Embedding Performance

Following the ImageNet-pretrained reference evaluation, DenseNet121, EfficientNet-B2, and MobileNetV3-Small were adapted to the LDHF-DB dataset to obtain dataset-specific embeddings. Their Rank-1 identification results for Minkowski orders $p = 1$ to 7 are summarized in the domain-adapted (DA) rows of Table 2. DenseNet121 achieved Rank-1 accuracies from 91.76% at $p = 1$ to a maximum of

93.19% at $p = 5$, representing a 1.43 percentage-point improvement. MobileNetV3-Small similarly increased from 88.17% at $p = 1$ to 91.40% at $p = 5$, an improvement of 3.23 percentage points, before decreasing at higher orders. EfficientNet-B2 showed a gradual increase from 79.57% at $p = 1$ to 84.23% at $p = 7$, corresponding to a 4.66-percentage-point improvement. These results demonstrate that increasing the Minkowski order generally improved Rank-1 identification, although the optimal order differed across architectures. The variation in Rank-1 performance is illustrated in Fig. 3.

The corresponding verification results are presented in the DA rows of Table 3. DenseNet121 achieved its lowest EER of 2.16% at $p = 5$, decreasing from 2.75% at $p = 1$. MobileNetV3-Small reached its lowest EER of 2.85% at $p = 6$, compared with 3.94% at $p = 1$, while EfficientNet-B2 achieved its lowest EER of 4.30% at $p = 6$, compared with 5.52% at $p = 1$. EER therefore generally decreased with increasing Minkowski order before reaching a minimum and slightly increasing at higher orders. Notably, the EER-optimal order did not match the Rank-1-optimal order for EfficientNet-B2 and MobileNetV3-Small.

C. Identification Performance Analysis

Identification performance in this subsection focuses

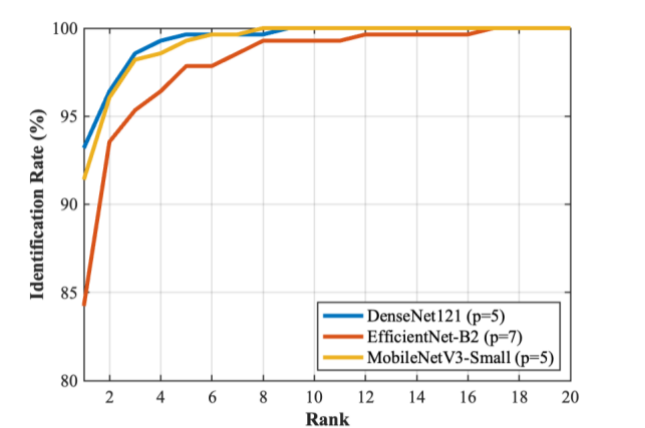


Fig. 4. CMC curves of domain-adapted CNN models using their respective Rank-1-optimal Minkowski distance orders.

distance combinations using the Rank-1-optimal Minkowski order for each model. The resulting Rank-1 accuracies are summarized in Table 4. For DenseNet121, accuracy ranged from 81.72% to 91.40%, with the highest performance observed for the 1 m-to-1 m pair and the lowest for the 100 m-to-60 m pair. EfficientNet-B2 achieved lower performance overall, ranging from 66.67% to 76.34%, with the highest accuracy for the 1 m-to-1 m and 1 m-to-100 m

Table 3. Equal error rate (EER) of pretrained (PT) and domain-adapted (DA) CNN models at different Minkowski distance orders (%).

Model	Type	p = 1	p = 2	p = 3	p = 4	p = 5	p=6	p=7
ResNet50	PT	47.38	45.74	45.80	46.92	46.84	46.82	46.90
DenseNet121	PT	45.63	45.62	46.11	46.48	46.36	46.48	46.59
EfficientNet-B2	PT	41.71	41.94	42.42	43.13	43.25	43.7	44.09
MobileNetV3-Small	PT	47.17	46.48	46.72	47.67	48.02	47.94	47.91
DenseNet121	DA	2.75	2.63	2.37	2.27	2.16	2.25	2.27
EfficientNet-B2	DA	5.52	5.25	4.78	4.51	4.41	4.30	4.31
MobileNetV3-small	DA	3.94	3.56	3.21	2.99	2.86	2.85	2.99

on the domain-adapted models, while the ImageNet-pretrained models were used as reference baselines. Identification performance was first assessed using cumulative match characteristic (CMC) curves, which show the probability that the correct identity is retrieved within the top-(k) ranked gallery candidates. Fig. 4 presents the CMC curves obtained using the Rank-1-optimal Minkowski order for each model. At Rank-1, DenseNet121 achieved the highest accuracy at 93.19%, followed by MobileNetV3-Small at 91.40% and EfficientNet-B2 at 84.23%. DenseNet121 reached 100% cumulative recognition by Rank-9, while MobileNetV3-Small and EfficientNet-B2 reached 100% by Rank-8 and Rank-17, respectively, indicating that DenseNet121 provided the strongest identification performance across the evaluated ranks.

Distance-specific identification performance was subsequently evaluated across all nine probe-gallery

pairs and the lowest for the 60 m-to-60 m pair. MobileNetV3-Small ranged from 74.19% to 90.32%, with its highest accuracy on the 60 m-to-60 m pair and its lowest on the 100 m-to-1 m pair. Across all models, the highest individual distance-pair accuracy was 91.40% for DenseNet121 at 1 m-to-1 m, whereas the lowest was 66.67% for EfficientNet-B2 at 60 m-to-60 m.

To provide an overall comparison, the distance-specific results in Table 4 were further summarized using Same-Distance Accuracy (SDA) and Cross-Distance Accuracy (CDA), following the procedure described in Section III-E. The resulting averages are presented in Table 5. DenseNet121 achieved an SDA of 88.89% and a CDA of 86.38%, corresponding to a 2.51 percentage-point difference. MobileNetV3-Small achieved an SDA of 86.38% and a CDA of 82.62%, giving a 3.76 percentage-point difference. In contrast, EfficientNet-B2 achieved an SDA of 71.68% and a CDA

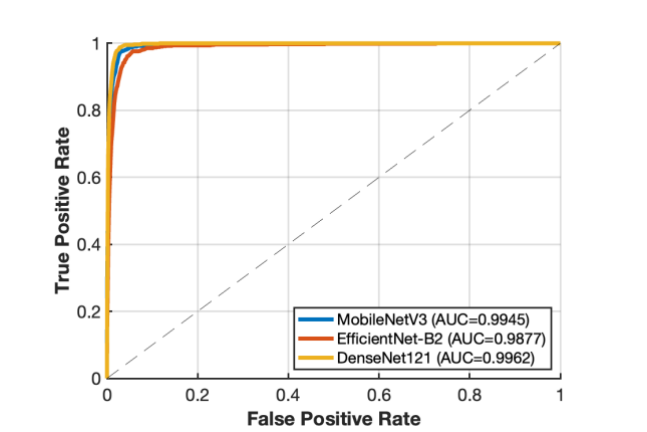


Fig. 5. ROC curves for the domain-adapted CNN models at their respective Rank-1-optimal Minkowski distance orders.

of 73.66%, with CDA exceeding SDA by 1.98 percentage points. These results show that the relationship between same-distance and cross-distance performance was architecture-dependent rather than uniformly favoring same-distance matching.

D. Verification Performance Analysis

Verification performance was evaluated using the Minkowski distance order that yielded the highest Rank-1 identification accuracy for each domain-adapted model, although the lowest EER is observed at a different order. Fig.5 presents the ROC curves, while Table 6 summarizes the verification results in

Table 4. Same-distance and cross-distance Rank-1 identification accuracy (%) of domain-adapted models using the Rank-1-optimal Minkowski distance order for each architecture.

Probe to Gallery	DenseNet 121 (p=5)	EfficientNet -B2 (p=7)	MobileNetV3 -Small (p=5)
1 to 1 m	91.40	76.34	83.87
1 to 60 m	86.02	75.27	86.02
1 to 100 m	83.87	76.34	80.65
60 to 1 m	90.32	67.74	82.80
60 to 60 m	87.10	66.67	90.32
60 to 100 m	90.32	75.27	83.87
100 to 1 m	86.02	73.12	74.19
100 to 60 m	81.72	74.19	88.17
100 to 100 m	88.17	72.04	84.95

terms of area under the curve (AUC), true acceptance rate (TAR) at FAR = 1% and FAR = 0.1%, and equal error rate (EER). DenseNet121 achieved the strongest verification performance, with an AUC of 0.9962, an EER of 2.16%, and TAR values of 89.73% at FAR = 1% and 44.21% at FAR = 0.1%. MobileNetV3-Small achieved an AUC of 0.9945, an EER of 2.86%, and TAR values of 85.42% and 43.01% at FAR = 1% and 0.1%, respectively. EfficientNet-B2 achieved an AUC of

0.9877, an EER of 4.31%, and TAR values of 70.97% and 23.06% at the corresponding FAR levels. Thus, DenseNet121 provided the best verification performance across all reported metrics, followed by MobileNetV3-Small and EfficientNet-B2.

E. Statistical Analysis

To determine whether the observed differences between the Euclidean baseline and the Rank-1-optimal Minkowski distance orders were statistically significant, confidence intervals and paired hypothesis tests were applied to both identification and verification performance. Table 7 compares ($p = 2$) with the distance order yielding the highest Rank-1 accuracy for each domain-adapted CNN. Rank-1 differences were assessed using the exact McNemar test, while paired differences in EER were evaluated using bootstrap resampling with 5,000 replicates and 95% confidence intervals.

For DenseNet121, increasing the Minkowski order from $p = 2$ to $p = 5$ increased Rank-1 accuracy from 92.47% to 93.19%. This improvement of 0.72 percentage points was not statistically significant according to McNemar's test ($p = 0.6250$). In contrast, EER decreased from 2.63% to 2.16%, corresponding to a reduction of 0.46 percentage points (95% CI: -0.60 to -0.08), which was statistically significant according to the paired bootstrap analysis ($p = 0.0164$).

Table 5. Same-Distance Accuracy (SDA) and Cross-Distance Accuracy (CDA) at the Rank-1-optimal Minkowski order for each model.

Model	SDA (%)	CDA (%)
DenseNet121 $p = 5$	88.89	86.38
MobileNetV3-Small $p = 5$	86.38	82.62
EfficientNet-B2 $p = 7$	71.68	73.66

For EfficientNet-B2, increasing the Minkowski order from $p = 2$ to $p = 7$ increased Rank-1 accuracy from 80.29% to 84.23%, an improvement of 3.94 percentage points that was statistically significant according to McNemar's test ($p = 0.0127$). EER decreased from 5.25% to 4.31%, corresponding to a reduction of 0.95 percentage points (95% CI: -1.34 to -0.36), which was also statistically significant ($p < 0.0004$). For MobileNetV3-Small, increasing the Minkowski order from $p = 2$ to $p = 5$ increased Rank-1 accuracy from 88.17% to 91.40%, an improvement of 3.23 percentage points that was statistically significant according to McNemar's test ($p = 0.0225$). EER decreased from 3.56% to 2.86%, corresponding to a reduction of 0.71 percentage points (95% CI: -0.94 to -0.20), which was also statistically significant according to the bootstrap analysis ($p = 0.0076$). Across the three domain-adapted CNNs, the bootstrap analysis showed statistically significant reductions in EER relative to the Euclidean baseline. McNemar's test indicated statistically significant Rank-1 improvements for

EfficientNet-B2 and MobileNetV3-Small, whereas the Rank-1 improvement for DenseNet121 was not statistically significant.

IV. Discussion

A. Pre-trained and Domain-Adapted Embeddings

This study investigated how feature embeddings and the generalized Minkowski distance jointly influence CSCD-FR performance. The consistently low Rank-1 accuracies obtained by the ImageNet-pretrained CNN models indicate that generic pretrained embeddings alone were insufficient under the evaluated conditions. Unlike conventional face recognition benchmarks, the LDHF-DB combines spectral differences between VIS and NIR images with variation in acquisition distance, resulting in substantially greater appearance variability. Since the pretrained networks were originally optimized for large-scale object recognition rather than cross-spectral face recognition, the extracted embeddings provided limited identity discrimination under these challenging conditions. Similar observations have been reported in transfer learning studies, where ImageNet-pretrained features require adaptation to the target domain to achieve competitive performance on specialized vision tasks [37][38][39].

After domain adaptation, all three CNN models demonstrated substantial improvements in both identification and verification performance. This improvement is consistent with previous studies showing that adapting pretrained CNNs to the target domain can produce more discriminative feature embeddings than directly using ImageNet-pretrained representations [17][39]. Optimizing the embedding layer using triplet margin loss encourages samples from the same identity to form more compact representations while separating samples from different identities, resulting in a feature space better suited for similarity-based matching.

B. Impact of Minkowski Distance on CNN Embedding

The experiments evaluated generalized Minkowski distance orders from $p = 1$ to $p = 7$ for each domain-adapted CNN model using both identification and verification metrics. Compared with the conventional Manhattan ($p = 1$) and Euclidean ($p = 2$) cases, higher Minkowski orders generally produced improved recognition performance, although the magnitude and statistical significance of these improvements varied across architectures. DenseNet121 and MobileNetV3-Small achieved their highest Rank-1 accuracies at $p = 5$, whereas EfficientNet-B2 performed best at $p = 7$. Statistical analysis further showed that the reduction in EER was significant for all three models, while statistically significant improvements in Rank-1 accuracy were observed only for EfficientNet-B2 and MobileNetV3-Small.

Although all feature vectors were L_2 -normalized, the three CNN architectures learned different embedding representations after domain adaptation. Consequently, each embedding space responded differently to changes in the Minkowski distance order. This observation is particularly relevant because deep face recognition systems commonly perform matching in a learned embedding space using Euclidean distance or cosine similarity [21][22]. The present findings suggest that similarity metric selection should be considered together with the learned embedding representation rather than assuming a single distance metric is optimal for all CNN architectures.

From a mathematical perspective, increasing the Minkowski distance order progressively increases the relative contribution of embedding dimensions with larger absolute differences. Thus, higher orders place greater emphasis on the largest coordinate-wise discrepancies between feature vectors. This behaviour provides a plausible explanation for why DenseNet121

Table 6. Verification performance of domain-adapted CNN models at their respective Rank-1-optimal Minkowski distance order.

Model	Minkowski order p	EER (%)	AUC	TAR @ FAR 1%	TAR @ FAR 0.1%
DenseNet121	5	2.16	0.9962	89.73	44.21
EfficientNet-B2	7	4.31	0.9877	70.97	23.06
MobileNetV3-Small	5	2.86	0.9945	85.42	43.01

Table 7. Statistical comparison between Euclidean distance and the Minkowski distance order yielding the highest Rank-1 accuracy for each domain-adapted model.

Models	Minkowski Order p	Rank-1 Accuracy (%) [95% CI]	McNemar p-value	EER (%) [95% CI]	Δ EER (%) [95% CI]	Bootstrap p-value
DenseNet121	$p = 2$	92.47 [89.25 95.34]	0.6250	2.63 [2.17 2.99]	-0.46 [-0.60 -0.08]	0.0164
	$p = 5$	93.19 [89.96 96.06]		2.16 [1.85 2.69]		
EfficientNet-B2	$p = 2$	80.29 [75.63 84.60]	0.0127	5.25 [4.43 5.90]	-0.95 [-1.34 -0.36]	<0.0004
	$p = 7$	84.23 [79.93 88.17]		4.31 [3.79 5.02]		
MobileNetV3-Small	$p = 2$	88.17 [84.23 91.76]	0.0225	3.56 [3.00 3.96]	-0.71 [-0.94 -0.20]	0.0076
	$p = 5$	91.40 [88.16 94.62]		2.86 [2.45 3.58]		

and MobileNetV3-Small reached their best performance at $p = 5$, while EfficientNet-B2 benefited from the stronger emphasis provided by $p = 7$. This observation is consistent with the broader findings of Aggarwal *et al.* [29], who showed that the effectiveness of Minkowski distance measures depends on the characteristics of high-dimensional data. While their analysis considered generic high-dimensional spaces, the present study extends this perspective to learned CNN embedding spaces optimized for CSCD-FR. This suggests that the

Correlation (BLPOC). These studies used image restoration, handcrafted features, or phase-based matching rather than the deep embedding and metric-learning framework adopted in the present study. More recent CSCD studies have reported high identification rates under different experimental conditions. Ai *et al.* al [17] combined an improved CycleGAN with edge loss for NIR–VIS translation and a FaceNet-based CMFRNet using MobileNet-v1 as the backbone feature-extraction network, reporting Rank-1 accuracies of 100% at 1 m and 90.59% at 8 m. The

Table 8. Comparison with previous Cross-Spectral Cross-Distance Face Recognition studies

Authors	Methods	Output
Kang et al. (2014) [10]	AHFR with NIR image restoration and HFR matching	EER: 8.6%; GAR: 73.7% at FAR = 1%
Shamia et al. (2017) [11]	DoG filtering; HOG, LBP, and normal-fitting features	Recognition rate: 72% (60 m), 78% (100 m), 32% (150 m)
Maeng et al. (2013) [12]	SIFT patch descriptors and learned discriminant projections	Verification rate: 81% (60 m), 61% (100 m), 20% (150 m) at FAR = 1%
Arnia et al. (2021) [13]	Homomorphic filtering and BLPOC	EER: 5.34%; GAR: 93% at FAR = 1%
Ai et. al (2024) [16]	Asymmetric CycleGAN	Rank-1: 90.00%; verification rate: 87.93% at FAR = 1%
Ai et al. (2026) [17]	Improved CycleGAN with edge loss; FaceNet-based CMFRNet with MobileNet-v1 backbone	Rank-1: 100% (1 m); 90.59% (8 m)
Present study	Domain-adapted CNN embeddings and Minkowski distance	DenseNet121: Rank-1 = 93.19% ($p = 5$); EER $\approx$ 2.16%

preferred Minkowski distance order may depend on learned embedding-space properties beyond dimensionality alone. Because the contribution of individual embedding dimensions was not explicitly analyzed, this interpretation should be regarded as a plausible explanation rather than a definitive mechanism.

C. Comparison with Previous CSCD-FR Studies

Table 8 summarizes previous CSCD-FR studies in terms of their methods, evaluation conditions, and reported performance. Early studies using LDHF-DB demonstrated the difficulty of long-distance NIR-to-VIS recognition. Kang *et al.* [10] reported an EER of 8.6% and a GAR of 73.7% at FAR = 1% using an image-restoration and heterogeneous face recognition approach. Shamia *et al.* [11] reported recognition rates of 72%, 78%, and 32% at 60, 100, and 150 m, respectively, using handcrafted feature descriptors. Maeng *et al.* al [12] reported verification rates of 81%, 61%, and 20% at FAR = 1% for 60, 100, and 150 m NIR probes, respectively, using SIFT-based local descriptors and learned discriminant projections. Arnia *et al.* [13] reported an EER of 5.34% and a GAR of 93% at FAR = 1% for their evaluated LDHF-DB cross-spectral and cross-distance condition using homomorphic filtering and Band-Limited Phase-Only

related asymmetric CycleGAN framework in [16] achieved 90.00% Rank-1 at 8 m and a verification rate of 87.93% at FAR = 1%. However, these studies used substantially shorter acquisition distances than the 60–100 m conditions evaluated here. Under the present LDHF-DB protocol, DenseNet121 achieved its highest observed Rank-1 accuracy of 93.19% at $p = 5$, with an EER of approximately 2.16%. This Rank-1 accuracy is higher than the 90.59% reported at 8 m in [17], although the experimental conditions differ substantially. The lower EER than those reported in [10] and [13] indicates favorable verification performance under the present experimental setting. However, we cannot claim direct superiority because of differences in dataset configuration, training strategy, distance conditions, and evaluation protocol.

D. Statistical Analysis

The statistical analysis shows that the effect of Minkowski-order selection depends on both the CNN architecture and the evaluation metric. Using the conventional Euclidean distance ($p = 2$) as the reference condition, the Rank-1-optimal Minkowski order produced statistically significant reductions in EER for all three domain-adapted CNN architectures. The EER reduction was 0.46 percentage points for DenseNet121, 0.95 percentage points for EfficientNet-

B2, and 0.71 percentage points for MobileNetV3-Small. The magnitude of the reduction therefore differed between architectures.

A different pattern was observed for closed-set identification. Rank-1 accuracy increased by 0.72 percentage points for DenseNet121, 3.94 percentage points for EfficientNet-B2, and 3.23 percentage points for MobileNetV3-Small relative to the $p = 2$ reference. The increases for EfficientNet-B2 and MobileNetV3-Small were statistically significant, whereas the increase for DenseNet121 was not. Nevertheless, DenseNet121 achieved the highest observed Rank-1 accuracy and the lowest EER of 2.16% among the evaluated models. Its 0.72-percentage-point improvement at $p = 5$ was therefore insufficient to establish a statistically significant difference from $p = 2$.

The different statistical outcomes between Rank-1 accuracy and EER are particularly relevant to CSCD-FR evaluation. Rank-1 accuracy evaluates whether the correct identity is retrieved as the closest gallery match, whereas EER evaluates the separation between genuine and impostor comparisons over a range of decision thresholds. A change in the distance metric may therefore have a limited effect on the nearest-neighbor decision while still altering the distributions of genuine and impostor distances. This distinction explains why the verification results provide additional information beyond Rank-1 accuracy. Overall, the statistical results indicate the potential value of Minkowski-order selection as an additional component of the CSCD-FR matching stage, while also showing that its effect is not uniform across architectures or evaluation criteria. Therefore, reporting both identification and verification measures is important when evaluating distance metrics.

E. Analysis of Distance-Specific Recognition

Distance-specific evaluation showed that recognition performance was not determined solely by matching probe and gallery acquisition distances. Although same-distance pairs generally performed well, some cross-distance combinations achieved higher Rank-1 accuracy. For example, DenseNet121 achieved 90.32% for the 60 m–100 m combination, compared with 87.10% for 60 m–60 m, while EfficientNet-B2 achieved 75.27% compared with 66.67%, respectively. These results indicate that acquisition distance alone does not determine recognition performance.

This behavior may reflect the interaction between the learned embedding representation and differences in image quality across acquisition conditions [40]. In particular, images acquired at the same nominal distance may not provide equally discriminative representations after preprocessing and embedding extraction. Consequently, a 100 m VIS gallery image may in some cases produce a more compatible embedding with a 60 m NIR probe than a 60 m VIS

gallery image. The 3×3 identification matrices show that certain gallery distances produced higher recognition accuracy across multiple probe distances. Thus, CDA may occasionally approach or exceed SDA because it includes several high-performing probe-gallery combinations, rather than because cross-distance matching is inherently superior. These findings support evaluating all probe-gallery distance combinations rather than relying only on same-distance matching.

F. Limitation, Future Work, and Practical Implication

The evaluation used the LDHF-DB dataset and three CNN architectures, with domain adaptation limited to the embedding layer while the convolutional backbones remained frozen. Because the same retained identities were used for adaptation and evaluation, the results represent a dataset-specific assessment rather than generalization to unseen subjects. In addition, the 150 m subset was excluded, limiting the conclusions to the 1, 60, and 100 m conditions. Future work should therefore evaluate subject-disjoint protocols, external datasets, and the excluded 150 m condition, while exploring additional CNN architectures and alternative similarity measures. The architecture-dependent optimal orders have practical implications for CSCD-FR system design. DenseNet121 and MobileNetV3-Small achieved their highest observed Rank-1 accuracy at $p = 5$, whereas EfficientNet-B2 performed best at $p = 7$. Since Minkowski-order selection occurs after feature extraction, it offers a relatively low-cost way to modify the matching stage without retraining the CNN backbone. Thus, optimal order should be determined empirically for the selected architecture and application rather than assumed to be universally optimal.

V. Conclusion

This study investigated the influence of generalized Minkowski distance order on deep embedding-based cross-spectral cross-distance face recognition (CSCD-FR). Using the LDHF-DB dataset, DenseNet121, EfficientNet-B2, and MobileNetV3-Small were evaluated under ImageNet-pretrained and domain-adapted settings across different Minkowski distance orders. The results showed that the optimal distance order was architecture-dependent. DenseNet121 achieved the highest observed Rank-1 accuracy of 93.19% and the lowest EER of 2.16% at $p = 5$. Distance-specific evaluation further showed that cross-distance matching could achieve comparable or higher performance than same-distance matching under several conditions. These findings demonstrate that distance-metric selection provides an additional optimization option at the matching stage without modifying the CNN backbone and can therefore be considered alongside the learned embedding

representation. Since the same retained identities were used for domain adaptation and evaluation, the results represent a dataset-specific empirical assessment rather than recognition performance on previously unseen subjects.

Future research should assess the generalizability of these findings using subject-disjoint protocols or nested cross-validation and larger, more diverse cross-spectral face datasets, including the excluded 150 m condition. Further investigation of alternative CNN architectures, domain adaptation strategies, metric-learning approaches, and similarity measures may provide additional insight into the relationship between learned embedding representations and similarity metric selection.

Declarations

Ethics Approval and Consent to Participate

This study used the publicly available Long-Distance Heterogeneous Face Database (LDHF-DB) and did not involve direct recruitment, intervention, or collection of new data from human participants by the authors. Therefore, additional institutional ethical approval and informed consent were not required for the present secondary-data analysis. The use of the dataset followed the conditions specified by the original data provider.

Consent for Publication

Not applicable. No new identifiable participant information was collected or reported in this study beyond data already available in the publicly released dataset.

Competing Interests

The authors declare that they have no competing financial or non-financial interests that could have influenced the work reported in this manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The data analyzed in this study were obtained from the publicly available Long-Distance Heterogeneous Face Database (LDHF-DB). No new dataset was generated during this study. The dataset can be accessed subject to the terms and conditions established by its original provider.

Code Availability

The computational code used for preprocessing, feature extraction, Minkowski-distance evaluation, and

statistical analysis is available from the corresponding author upon reasonable request.

Author Contribution

Ernita Dewi Meutia: Conceptualization, Methodology, Investigation, Software, Data Curation, Formal Analysis, Visualization, Writing – Original Draft.

Fitri Arnia: Conceptualization, Methodology, Supervision, Validation, Formal Analysis, Writing – Review & Editing.

Nasaruddin: Supervision, Validation, Writing – Review & Editing.

Rusdha Muharar: Conceptualization, Methodology, Formal Analysis, Interpretation, Writing – Review & Editing.

All authors reviewed and approved the final manuscript and agreed to be accountable for the integrity of the work.

Acknowledgments

The authors acknowledge the developers and providers of the Long-Distance Heterogeneous Face Database (LDHF-DB) for making the dataset available for academic research. If there was no additional technical/institutional assistance, this statement is sufficient.

Use of Generative Artificial Intelligence

During the preparation of this manuscript, the authors used generative artificial intelligence tools to assist with language improvement, paraphrasing, code development/debugging, literature-search support, and manuscript-quality checking. The authors critically reviewed and verified all AI-assisted outputs and take full responsibility for the accuracy, integrity, originality, interpretation, and conclusions of the manuscript. Generative AI tools were not listed as authors and were not used as substitutes for author responsibility.

References

- [1] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face Recognition Systems: A Survey," *Sensors*, vol. 20, no. 2, p. 342, Jan. 2020, doi: [10.3390/s20020342](https://doi.org/10.3390/s20020342).
- [2] J. Soldera and J. Scharcanski, "Face Recognition, Modalities and Applications [Instrumentation and Measurement Systems]," *IEEE Instrum. Meas. Mag.*, vol. 29, no. 1, pp. 13–20, Feb. 2026, doi: [10.1109/MIM.2026.11395469](https://doi.org/10.1109/MIM.2026.11395469).
- [3] S. Jain, V. Chaudhari, R. Chuadhari, T. Chavan, and P. Shahane, "A Survey on Face Recognition Techniques in Machine Learning," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 8, no. 6, pp. 50–66, Nov. 2022, doi: [10.32628/cseit228558](https://doi.org/10.32628/cseit228558).

- [4] D. Anghelone, C. Chen, A. Ross, and A. Dantcheva, "Beyond the visible: A survey on cross-spectral face recognition," *Neurocomputing*, vol. 611, p. 128626, Jan. 2025, doi: [10.1016/j.neucom.2024.128626](https://doi.org/10.1016/j.neucom.2024.128626).
- [5] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Cross-spectral Face Completion for NIR-VIS Heterogeneous Face Recognition," Feb. 10, 2019, *arXiv*: [arXiv:1902.03565](https://arxiv.org/abs/1902.03565). doi: [10.48550/arXiv.1902.03565](https://doi.org/10.48550/arXiv.1902.03565).
- [6] S. Wahyudi, M. Oktiana, F. Arnia, and K. Munadi, "Cross-Spectral Face Matching Based on Phase Correlation," in *Proc. 2019 IEEE Int. Conf. Cybernetics Comput. Intell. (CyberneticsCom)*, Aug. 2019, pp. 23–28, doi: [10.1109/CYBERNETICSCOM.2019.8875642](https://doi.org/10.1109/CYBERNETICSCOM.2019.8875642).
- [7] E. D. Meutia, F. Arnia, Nasaruddin, R. Muharar, M. Oktiana, and K. Saddami, "Performance Evaluation of Deep Feature in Cross Spectral Face Recognition," in *Proc. 2024 11th Int. Conf. Inf. Technol., Comput., Electr. Eng. (ICITACEE)*, Aug. 2024, pp. 314–319, doi: [10.1109/ICITACEE62763.2024.10762811](https://doi.org/10.1109/ICITACEE62763.2024.10762811).
- [8] N. D. Kalka, T. Bourlai, B. Cukic, and L. Hornak, "Cross-spectral face recognition in heterogeneous environments: A case study on matching visible to short-wave infrared imagery," in *Proc. 2011 Int. Joint Conf. Biometrics (IJB)*, Oct. 2011, pp. 1–8, doi: [10.1109/IJB.2011.6117586](https://doi.org/10.1109/IJB.2011.6117586).
- [9] I. N. Jarudi *et al.*, "Recognizing distant faces," *Vision Res.*, vol. 205, p. 108184, Apr. 2023, doi: [10.1016/j.visres.2023.108184](https://doi.org/10.1016/j.visres.2023.108184).
- [10] D. Kang, H. Han, A. K. Jain, and S.-W. Lee, "Nighttime face recognition at large standoff: Cross-distance and cross-spectral matching," *Pattern Recognit.*, vol. 47, no. 12, pp. 3750–3766, Dec. 2014, doi: [10.1016/j.patcog.2014.06.004](https://doi.org/10.1016/j.patcog.2014.06.004).
- [11] D. Shamia and D. A. Chandy, "Intelligent system for cross-spectral and cross-distance face matching," *Comput. Electr. Eng.*, vol. 71, pp. 915–924, Oct. 2018, doi: [10.1016/j.compeleceng.2017.09.004](https://doi.org/10.1016/j.compeleceng.2017.09.004).
- [12] H. Maeng, S. Liao, D. Kang, S.-W. Lee, and A. K. Jain, "Nighttime Face Recognition at Long Distance: Cross-Distance and Cross-Spectral Matching," in *Computer Vision – ACCV 2012* (Lecture Notes in Computer Science, vol. 7725), K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, Eds. Berlin, Heidelberg: Springer, 2013, pp. 708–721, doi: [10.1007/978-3-642-37444-9_55](https://doi.org/10.1007/978-3-642-37444-9_55).
- [13] F. Arnia, M. Oktiana, K. Saddami, K. Munadi, R. Roslidar, and B. Pradhan, "Homomorphic Filtering and Phase-Based Matching for Cross-Spectral Cross-Distance Face Recognition," *Sensors*, vol. 21, no. 13, p. 4575, Jul. 2021, doi: [10.3390/s21134575](https://doi.org/10.3390/s21134575).
- [14] W. Hu, Y. Yang, and H. Hu, "Pseudo Label Association and Prototype-Based Invariant Learning for Semi-Supervised NIR-VIS Face Recognition," *IEEE Trans. Image Process.*, vol. 33, pp. 1448–1463, 2024, doi: [10.1109/TIP.2024.3364530](https://doi.org/10.1109/TIP.2024.3364530).
- [15] M. Tarasiou, J. Deng, and S. Zafeiriou, "Rethinking the Domain Gap in Near-infrared Face Recognition," in *Proc. 2024 IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 940–949, doi: [10.1109/CVPRW63382.2024.00099](https://doi.org/10.1109/CVPRW63382.2024.00099).
- [16] D. Ai, K. Jia, Y. Wang, and Y. Liu, "NIR-VIS Image Translation for the Cross-Spectral and Cross-Distance Face Recognition," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, Jul. 2024, pp. 1–6. doi: [10.1109/ICME57554.2024.10687584](https://doi.org/10.1109/ICME57554.2024.10687584).
- [17] D. Ai, Y. Wang, K. Jia, Z. Ji, and Y. Liu, "Cross-distance near-infrared face recognition," *J. Vis. Commun. Image Represent.*, vol. 115, p. 104691, Jan. 2026, doi: [10.1016/j.jvcir.2025.104691](https://doi.org/10.1016/j.jvcir.2025.104691).
- [18] E. Vezzetti, *Similarity Measures for Face Recognition*. Oak Park: Bentham Science Publishers, 2015, doi: [10.2174/97816810804441150101](https://doi.org/10.2174/97816810804441150101).
- [19] S. S. Rajput, C. Yadav, and U. Chourasia, "Analysis of Euclidean Distance, Manhattan Distance, and Cosine Similarity for Image Retrieval Through Deep Learning," in *Proc. 2025 Int. Conf. Eng. Innovations Technol. (ICoEIT)*, Jul. 2025, pp. 1310–1315, doi: [10.1109/ICoEIT63558.2025.11211678](https://doi.org/10.1109/ICoEIT63558.2025.11211678).
- [20] G. Qian, S. Sural, Y. Gu, and S. Pramanik, "Similarity between Euclidean and cosine angle distance for nearest neighbor queries," in *Proc. 2004 ACM Symp. Applied Computing*, Mar. 2004, pp. 1232–1237, doi: [10.1145/967900.968151](https://doi.org/10.1145/967900.968151).
- [21] 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](https://doi.org/10.1109/TPAMI.2021.3087709).
- [22] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. 2015 IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823, doi: [10.1109/CVPR.2015.7298682](https://doi.org/10.1109/CVPR.2015.7298682).
- [23] C. C. Aggarwal, *Similarity and Distances*, in *Data Mining*. Cham, Switzerland: Springer,

- 2015, pp. 63–91, doi: [10.1007/978-3-319-14142-8_3](https://doi.org/10.1007/978-3-319-14142-8_3).
- [24] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," in *Database Theory — ICDT 2001* (Lecture Notes in Computer Science, vol. 1973), J. Van Den Bussche and V. Vianu, Eds. Berlin, Heidelberg: Springer, 2001, pp. 420–434, doi: [10.1007/3-540-44503-X_27](https://doi.org/10.1007/3-540-44503-X_27).
- [25] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. 2019 IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, Oct. 2019, pp. 1314–1324, doi: [10.1109/ICCV.2019.00140](https://doi.org/10.1109/ICCV.2019.00140).
- [26] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv preprint arXiv:1905.11946*, 2019, doi: [10.48550/ARXIV.1905.11946](https://doi.org/10.48550/ARXIV.1905.11946).
- [27] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proc. 2017 IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: [10.1109/CVPR.2017.243](https://doi.org/10.1109/CVPR.2017.243).
- [28] M. Ghosh, T. Sarkar, D. Chokhani, and A. Dey, "Face detection and extraction using Viola–Jones algorithm," in *Computational Advancement in Communication, Circuits and Systems*, Lecture Notes in Electrical Engineering, 2021, pp. 93–107, doi: [10.1007/978-981-16-4035-3_9](https://doi.org/10.1007/978-981-16-4035-3_9).
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. 2016 IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [30] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace: L2 Hypersphere Embedding for Face Verification," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1041–1049, doi: [10.1145/3123266.3123359](https://doi.org/10.1145/3123266.3123359).
- [31] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, Jan. 2004, doi: [10.1109/TCSVT.2003.818349](https://doi.org/10.1109/TCSVT.2003.818349).
- [32] R. Ranjan, A. Bansal, J. Zheng, H. Xu, J. Gleason, and B. Lu, "A Fast and Accurate System for Face Detection, Identification, and Verification," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 1, no. 2, pp. 82–96, Apr. 2019, doi: [10.1109/TBIOM.2019.2908436](https://doi.org/10.1109/TBIOM.2019.2908436).
- [33] L. L. Chambino, J. S. Silva, and A. Bernardino, "Multispectral Facial Recognition: A Review," *IEEE Access*, vol. 8, pp. 207871–207883, 2020, doi: [10.1109/ACCESS.2020.3037451](https://doi.org/10.1109/ACCESS.2020.3037451).
- [34] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000, doi: [10.1109/34.879790](https://doi.org/10.1109/34.879790).
- [35] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: [10.1007/BF02295996](https://doi.org/10.1007/BF02295996).
- [36] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman and Hall/CRC, 1994, doi: [10.1201/9780429246593](https://doi.org/10.1201/9780429246593).
- [37] L. L. Chambino, J. S. Silva, and A. Bernardino, "Multispectral Face Recognition Using Transfer Learning with Adaptation of Domain Specific Units," *Sensors*, vol. 21, no. 13, p. 4520, Jan. 2021, doi: [10.3390/s21134520](https://doi.org/10.3390/s21134520).
- [38] Z. H. Abbas and S. H. Shaker, "The Transfer Learning Models for Face Recognition: A Survey," in *2022 2nd International Conference on Advances in Engineering Science and Technology (AEST)*, Oct. 2022, pp. 764–768, doi: [10.1109/AEST55805.2022.10412930](https://doi.org/10.1109/AEST55805.2022.10412930).
- [39] G. Lokku, G. H. Reddy, and M. N. G. Prasad, "A Robust Face Recognition model using Deep Transfer Metric Learning built on AlexNet Convolutional Neural Network," in *2021 International Conference on Communication, Control and Information Sciences (ICCISc)*, Idukki, India: IEEE, Jun. 2021, pp. 1–6, doi: [10.1109/ICCISc52257.2021.9484935](https://doi.org/10.1109/ICCISc52257.2021.9484935).
- [40] N. Zhang, "A Study on the Impact of Face Image Quality on Face Recognition in the Wild," Jul. 05, 2023, *arXiv*: [arXiv:2307.02679](https://arxiv.org/abs/2307.02679), doi: [10.48550/arXiv.2307.02679](https://doi.org/10.48550/arXiv.2307.02679).

Author Biography



Ernita Dewi Meutia earned her Bachelor's degree in Electrical Engineering from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 1991. She obtained a Postgraduate Diploma in Electronics and Information Technology from the University of Birmingham, UK, in 1996. In 2009, she completed her Master's degree in Telecommunication Engineering at the University of Arkansas, USA, under a Fulbright Scholarship. She has been a faculty member at Universitas Syiah Kuala, Banda Aceh, Indonesia since 1992. She is currently a Doctoral degree student at Universitas Syiah Kuala. Her research interests include telecommunication engineering and deep learning.

Rusdha Muharar (Member, IEEE) received the B.Eng. degree in electrical engineering (cum laude) from Gadjah Mada University, Indonesia, in 1999, the M.Sc. degree in electrical engineering from TU Delft, The Netherlands, in 2004, and the Ph.D. degree in electrical



engineering from the University of Melbourne, Australia, in 2012. He joined the Department of Electrical Engineering at Universitas Syiah Kuala, Indonesia, in April 2006, where he is currently a professor. From November 2012 to November 2013, he was also a Postdoctoral Research

Fellow with the Department of Electrical and Computer Systems Engineering, Monash University, Australia. His research interests include communications theory, signal processing for wireless communications, and machine learning.



Nasaruddin received the B.Eng. degree in Electrical Engineering from Sepuluh Nopember Institute of Technology, Surabaya, Indonesia, in 1997. Then he received M. Eng. and D. Eng. in Physical Electronics and Informatics from the Graduate School of Engineering, Osaka

City University, Japan, in 2006 and 2009, respectively. He was head of The Master of Electrical Engineering Program, Graduate School of Syiah Kuala University. He is a professor in the Electrical and Computer Engineering Department, Syiah Kuala University. He has published several papers in international journals and accredited national journals. His research interests include digital communications, information theory, wireless communications, optical communications and ICT applications for disaster.



Fitri Arnia (Member, IEEE) received the B.Eng. degree from Universitas Sumatera Utara (USU), Medan, in 1997, the master's degree from the University of New South Wales (UNSW), Sydney, Australia, in 2004, and the Ph.D. degree from Tokyo Metropolitan University

(TMU), Tokyo, Japan, in 2008. She has been with the Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, since 1999. In November 2019, she became a Professor at Universitas Syiah Kuala. She was a Visiting Scholar with TMU in 2013 and Suleyman Demirel University (SDU), Isparta, Türkiye, in 2017. Her research interests include signal, image, and multimedia.