Minimizing Number of Distinct Poses for Pose-Invariant Face Recognition

Carter Ung, Pranav Mantini and Shishir K. Shah
Department of Computer Science, University of Houston, Houston, TX, U.S.A.
{ctung, sshah5, pmantini}@uh.edu

Keywords: Face Recognition, Computer Vision.

Abstract: In unconstrained environments, extreme pose variations of the face are a long-standing challenge for person identification systems. The natural occlusion of necessary facial landmarks is notable to model performance degradation in face recognition. Pose-invariant models are data-hungry and require large variations of pose in training data to achieve comparable accuracy in recognizing faces from extreme viewpoints. However, data collection is expensive and time-consuming, resulting in a scarcity of facial datasets with large pose variations for model training. In this study, we propose a training framework to enhance pose-invariant face recognition by identifying the minimum number of poses for training deep convolutional neural network (CNN) models, enabling higher accuracy with minimum cost for training data. We deploy ArcFace, a state-of-the-art recognition model, as a baseline to evaluate model performance in a probe-gallery matching task across groups of facial poses categorized by pitch and yaw Euler angles. We perform training and evaluation of ArcFace on varying pose bins to determine the rank-1 accuracy and observe how recognition accuracy is affected. Our findings reveal that: (i) a group of poses at -45°, 0°, and 45° yaw angles achieve uniform rank-1 accuracy across all yaw poses, (ii) recognition performance is better with negative pitch angles than positive pitch angles, and (iii) training with image augmentations like horizontal flips results in similar or better performance, further minimizing yaw poses to a frontal and $\frac{3}{4}$ view.

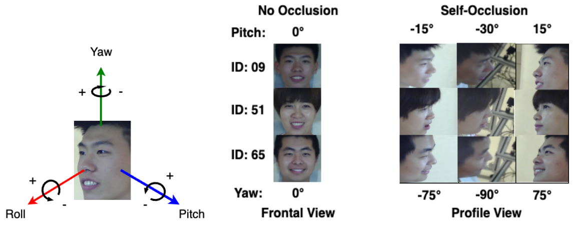
1 INTRODUCTION

Face recognition, widely known as a classic computer vision task, represents a long-standing research area that has grown in interest in recent years due to the developments of powerful deep convolutional neural networks (CNNs) (Taigman et al., 2014; Deng et al., 2019; Yin and Liu, 2018). While CNNs have proved to saturate accuracy in large face datasets, these feats are primarily restricted to images dominated by the frontal profile of the face. In real-world surveillance systems, captured faces can result in extreme orientations of the face due to the camera-to-subject perspective (Cheng et al., 2018). In these unconstrained scenarios where camera viewpoints and head orientation vary, the head pose of a person presents a natural self-occlusion of the face, leading to poor model performance for facial matching tasks (Zhang and Gao, 2009; Ahmed et al., 2019). Fig. 1 introduces the rotational angles that define head orientation and how extreme profile views lead to self-occlusion.

Several works have attempted to tackle pose-invariant challenges in face recognition (Asthana et al., 2011; Masi et al., 2016; Prince and Elder, 2006; Yin and Liu, 2018), however, the proposed models are often data-hungry and require diverse training data incorporating uniform distribution across pitch and yaw pose ranges to yield comparable results in facial matching tasks (Baltanas et al., 2021; Yin et al., 2019). As data collection presents a time-consuming and expensive process, the availability of meaningful training data for face recognition is limited (Chen et al., 2018). This is also true for the deployment of a trained face recognition system where rather limited poses are available for a subject to be enrolled in the gallery. Pivotal work has been explored in the M2FPA dataset introduced by Li *et al.* (Li et al., 2019), where the distribution of facial poses improves upon previous pose-aware benchmarks by providing abundant images across 66 pitch-yaw angles. However, their study did not evaluate unique combinations of poses for training and inference to analyze the complete model behavior across pitch and yaw. **We expand on their study using their M2FPA dataset to study model behavior in pose-occluded face recognition tasks.**

Ung, C., Mantini, P. and Shah, S. K.
Minimizing Number of Distinct Poses for Pose-Invariant Face Recognition.
DOI: 10.5220/0013186400003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 447-455
ISBN: 978-989-758-728-3; ISSN: 2184-4321



(a) Pitch, Yaw, Roll (b) Self-Occlusion Due to Pose

Figure 1: (a) Head pose is defined by three rotational Euler angles - pitch, yaw, and roll, (b) Faces captured at profile angles occlude the lateral side of the face, limiting discriminative features like eyes, cheek, and jawline (Li et al., 2019).

Addressing excessive data collection efforts, our aim is to identify a minimum number of poses needed in the gallery for each subject for uniform recognition across the face's 3D rotational plane. Building upon related works (Favelle and Palmisano, 2018; Bruce et al., 1987; Deng et al., 2019), our work is two-fold. First, we explore the empirical relationship between face recognition and pose discrepancies related to model performance. Here, we base our evaluation on experimental questions identifying (i) the minimal set of yaw angles in the gallery that ensures uniform recognition across all poses, (ii) the pitch angles combined with the optimal yaw set to optimize model performance, and (iii) whether synthetic pose augmentations can replace real poses while maintaining or improving recognition accuracy. We choose M2FPA (Li et al., 2019) for our extensive evaluation across the full natural range of yaw and pitch. We analyze face-matching accuracy during model inference from a set of facial poses containing unique pitch-yaw angles. Our key observations are generated by ablating key poses from current pose sets to gradually minimize the number of poses for the next pose set being evaluated. For example, given a set of angles between $-90°$ and $90°$, we decrement the number of poses in each iteration to validate the presence of degradation within the matching accuracy. Second, we propose a data selection protocol that determines the set of poses that are integrated into the training set and gallery for probe-gallery face matching. Following our pose ablation analysis, we optimize our minimized set of pitch-yaw angles by leveraging geometric pose augmentation techniques to replace mirrored faces with synthetic poses. Here, we achieve uniform recognition by filtering pose to only frontal and a $\frac{3}{4}$ view within training and gallery enrollment. We evaluate our results by fine-tuning pose sets against a pretrained ArcFace (Deng et al., 2019) model containing an iResNet-50 (Duta et al., 2021) CNN backbone. Our findings alleviate the current bottlenecks for data collection and provide guidance towards optimizing facial matching accuracy across any pose while minimizing required 2D poses per subject in the gallery enrollment, facilitating efficient processes for teaching CNN models pose-invariant capabilities in unconstrained camera environments. In summary, the contributions of this paper can be outlined as follows:

- We comprehensively analyze model performance behavior across pose ranges in pitch and yaw.
- We propose a training data selection protocol that minimizes the number of poses necessary for uniform recognition across all poses using distinct pose filtering and geometric augmentations.

2 RELATED WORKS

2.1 Pose-Aware Face Recognition

Pose is a long-standing factor for performance degradation for several identification tasks including face recognition (Zhang and Gao, 2009; Ahmed et al., 2019; Rajalakshmi and Jeyakumar, 2012), face detection (Qi et al., 2023; Zhang et al., 2016; Torres Pereira et al., 2014; Deng et al., 2020), and person re-identification (Nguyen et al., 2024; Khaldi et al., 2024). Early works prioritized holistic and classical machine learning techniques (Turk and Pentland, 1991; Belhumeur et al., 1997), producing accurate yet limited performance when exposed to facial changes and occlusion. However, state-of-the-art results have transitioned from traditional means to powerful CNNs that capture discriminative representations of the face, performing accurate recognition regardless of occluding factors (Yi et al., 2014; K. Wickrama Arachchilage and Izquierdo, 2020). The introduction of margin-based penalties inside the typical softmax loss function has proven effective in promoting inter-class separation and intra-class compactness. SphereFace (Liu et al., 2017), ArcFace (Deng et al., 2019), and CosFace (Wang et al., 2018) have leveraged margin penalty techniques to score high accuracy against notable unconstrained face benchmarks such as IJB-A (Klare et al., 2015), IJB-C (Maze et al., 2018), and LFW (Huang et al., 2007). Coupled with powerful feature extraction capabilities from novel CNN architectures (He et al., 2016; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014), CNN models are furthering their resiliency to challenging recognition factors including pose, illumination, and expression. However, it is apparent that benchmark results are dominated by evaluation on frontal faces, and further evaluation is needed on extreme view-point variations.

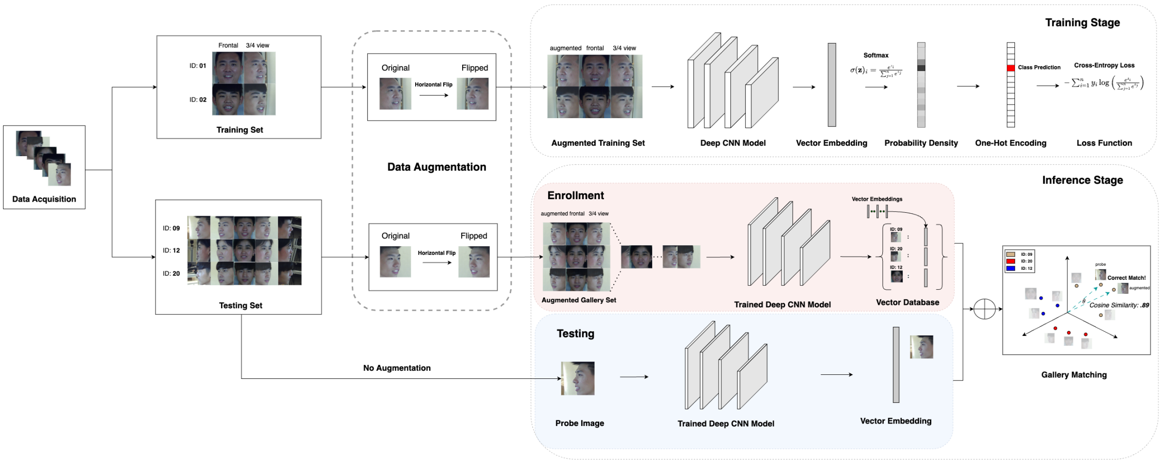


Figure 2: CNN model data selection framework composed of a training stage and inference stage pipeline. We enroll yaw poses 45° ($\frac{3}{4}$ view), 0° (frontal view), and synthetic mirrors of $\frac{3}{4}$ views in our training set and gallery, capable of identification regardless of yaw orientation.

2.2 Performance Across Pose

Studies focusing on model behavior across poses utilize model inference in face gallery matching tasks to observe recognition capabilities across the full head orientation (Gunawan and Prasetyo, 2017). Muller *et al.* (Müller et al., 2007) conducted a similarity rank-matching study that utilized statistical clustering to recognize -45°, 0°, and 45° yaw poses in a probe-gallery matching task. The study concluded that 0° yaw generalized the best across all three poses. These studies follow large motivation from natural recognition mechanisms in human psychology. It is believed that $\frac{3}{4}$ side views (45° yaw) of the face have been inferred to have an advantage over frontal and full side profile views when generalizing across facial views in psychological subject tests (Bruce et al., 1987). Favelle *et al.* (Favelle and Palmisano, 2018) conducted a psychological study of visual face recognition in nature. They investigated face recognition mechanisms with 80 human subjects participating in experiments recognizing faces from various viewpoints to understand the impacts of view-specific generalizations. The experimental protocol took a comparison view and matched it against a full range of pitch-yaw poses up to 75° in both rotation axes and deduced that $\frac{3}{4}$ view in yaw generalized best across views ranging from frontal to side profile orientations. Additionally, pitch comparison views looking down were favored in view-specific generalizations compared to extreme pitch angles where the face was projected upwards.

3 METHOD

3.1 Study Design Overview

As illustrated in Fig. 2, the proposed experimental framework is composed of several stages spanning both training and inference. First, we describe the process from data acquisition to the training pipeline. Given a set of N facial images $X = \{I_i \in \mathbb{R}^{512\times512}\}_{i=1}^{N}$ comprised of identity labels $Y = \{y_i \in \mathbb{R}\}_{i=1}^{N}$ obtained from data acquisition, we select a subset of identities $Y_{train} \subseteq Y$ for our training set $X_{\text{train}} \subseteq X$. The training set contains all poses for each identity in Y_{train}, and is filtered based on head pose, $p = \{\text{pitch}, \text{yaw}\}$, selecting specific pitch and yaw combinations that maximize performance. Let p^{pitch} denote the pitch angle and p^{yaw} denote the yaw angle. Given that X_{train} contains A pitch angles $K = \{p^{yaw} \in \mathbb{R}\}_{i=1}^{A}$ and M yaw angles $L = \{p^{pitch} \in \mathbb{R}\}_{i=1}^{M}$, we select a subset of pitch poses $K_{opt} \subseteq K$ and yaw poses $L_{opt} \subseteq L$ that yield optimal performance for facial matching. The training set after pose selection is then defined as:

$$X_{\text{train}} = \{I_i \in X \mid p_i^{\text{pitch}} \in K_{opt} \cap p_i^{\text{yaw}} \in L_{opt}\}.$$

X_{train} is fed into a deep CNN model for training or fine-tuning, where embeddings $f_i \in \mathbb{R}^{512}$ are generated to represent extracted discriminative features. The model is trained using standard loss functions such as softmax and cross-entropy to adjust embeddings in each epoch.

Second, we describe the protocol for model inference. We gather a set of identities $Y_{test} \subseteq Y$ for the testing set $X_{\text{test}} \subseteq X$, where each identity $y_i \in Y_{test}$ is distinct from those in the training set, i.e.,

$Y_{test} \cap Y_{train} = \emptyset$. Like the training set, X_{test} contains all poses for each identity. During inference, two sub-stages occur: enrollment and testing.

In the enrollment sub-stage, we filter the testing set X_{test} to retain only the poses that optimize performance based on the same criteria used for training. Specifically, we select the same optimal pose subset of pitch poses K_{opt} and yaw poses L_{opt}, and apply flip transformations to images where $p^{\text{yaw}} \neq 0°$, enrolling a pair of mirrored poses comprised of a non-augmented image and its flipped counterpart. The gallery set G, containing the known face identities, is then defined as:

$$G = \{I_i \in X_{\text{test}} \mid p_i^{\text{pitch}} \in K_{\text{opt}} \cap (p_i^{\text{yaw}} \in L_{\text{opt}} \cup \text{flip}(p_i^{\text{yaw}}))\}.$$

The faces in the gallery set are passed through the trained CNN to generate embeddings $f_G = \{f_g \in \mathbb{R}^{512} \mid g \in G\}$, which are stored in a vector database corresponding to their identity labels.

During the testing sub-stage, we select a probe image f_{probe} of unknown identity from X_{test}, where:

$$f_{\text{probe}} \in \{I_k \in X_{\text{test}} \mid p_k^{\text{pitch}} \in \mathbb{R} \cap p_k^{\text{yaw}} \in \mathbb{R}\}.$$

Here, the probe image is selected from all possible permutations of pitch and yaw angles within X_{test}. We note the probe set contains both optimal and non-optimal poses, indicating that the poses of the probe set include permutations of pitch and yaw angles that may not be in the gallery set. Following, we feed the probe image into the trained CNN to generate its embedding $f_{\text{probe}} \in \mathbb{R}^{512}$. To predict the probe image, we compute the cosine similarity between the probe embedding and each embedding in the gallery set where the gallery image $g_{\text{match}} = \arg\max_{g \in G} \cos(f_{\text{probe}}, f_g)$ corresponds to the highest cosine similarity, thus identifying the predicted matching identity for the probe image.

3.2 Pose Evaluation

Given a set of m poses, we utilize facial poses of pitch-yaw rotation angles $p_1, p_2, \ldots, p_m$, where each $p_i \in \mathbb{R}^2$, and group them in unique pose sets S_i, where $S_i \subseteq \{p_1, p_2, \ldots, p_m\}$ is a subset of the poses incorporated into the query and gallery sets during facial matching.

We organize images in the training and test sets into unique pose groups dependent on yaw, pitch, or augmentation evaluation. We use a dataset comprised of yaw and pitch angles, denoted as $Y = \{y_1, y_2, \ldots, y_m\}$ and $P = \{p_1, p_2, \ldots, p_m\}$, respectively, to create unique sets of pitch-yaw angles. Specifically, for a given task, we form a pose set $S_i \subseteq Y \times P$ where each S_i can include specific pitch and yaw combinations, such as grouping side profiles together, frontal poses, or combining extreme side profiles with frontal poses. Our baseline DCNN model is fine-tuned on the targeted pose group and then evaluated using a CNN backbone to encode face embeddings $f \in \mathbb{R}^{512}$ in the latent space for model inference. Let Q and G denote the query and gallery sets, respectively, each with distinct pose sets S_Q and S_G, such that S_Q and S_G are subsets of $Y \times P$. First, we convert all images in Q and G to embeddings. Then, we iterate through the query set, Q, comparing each probe image $q_i \in Q$ to the gallery of enrolled images G, computing the cosine similarity between each probe image q_i and each gallery image $g_j \in G$ separately. In a one-to-many comparison, we take the probe-gallery face pair (q_i, g_j) that computes the highest cosine similarity in the latent space as the model prediction.

Following pose evaluation methods to find a minimal pose set, we utilize augmentation techniques to synthesize mirrored poses to replace real faces. Given a set of poses S, we take a pose $p_i \in S$ where $p_i^{yaw} \neq 0$. We achieve the mirrored pose of p_i through the function $f(p_i) = \text{flip}(p_i)$, where *flip* represents a horizontal flip of the face image using geometric augmentation techniques. We incorporate the synthetic pose $p_{\text{fake}} = f(p_i)$ as a replacement for the real mirror pose $-p_i$, resulting in a new augmented pose set $S_{\text{aug}} = (S \setminus \{-p_i\}) \cup \{p_{\text{fake}}\}$. S_{aug} is incorporated into the gallery G and training set T for training and evaluation against a mixture of real and synthetic poses of faces. The augmentation effort aims to reduce the number of minimal poses m to $\lceil \frac{m}{2} \rceil$ yaw poses.

4 EXPERIMENTS

4.1 Experimental Outline

Our objective is to identify a minimal set of poses that ensures consistent recognition performance across varying head orientations. To achieve this, we designed a series of experiments aimed at validating our strategy for optimal data selection in facial matching tasks. The experiments, summarized in Table 1, focus on evaluating model performance across different yaw and pitch angles, as well as the impact of synthetic pose augmentation.

4.1.1 Datasets

The Multi-Pitch Multi-Yaw Dataset (M2FPA) (Li et al., 2019) contains 397,544 images of 229 subjects

Table 1: Summary of Experiments.

Experiment	Description
1. Yaw Evaluation	Find a minimal set of yaw angles that provides uniform recognition against all yaw poses.
2. Pitch Evaluation	Based on the optimal set of yaw angles determined in the first experiment, evaluate different pitch angles to identify the pitch angles that optimize model performance.
3. Pose Augmentation	Test performance with synthetic pose augmentation and checks whether eliminating mirrored poses can improve or maintain model performance compared to using real poses.
4. Gallery Reduction	Test performance with reduced gallery set where one image per yaw angle is enrolled for each ID.

with 62 poses (including 13 yaw angles, 5 pitch angles, and 44 yaw-pitch angles) which range from -90° to 90° degrees in yaw and -30° to 30° degrees in pitch. The distinct yaw poses comprise of $-90°$, $-75°$, $-67.5°$, $-60°$, $-45°$, $-30°$, $-22.5°$, $-15°$, 0°, 15°, 22.5°, 30°, 45°, 60°, 67.5°, 75°, and 90°. The distinct pitch poses are $-30°$, $-15°$, 0°, 15°, and 30°. Our study utilizes the M2FPA dataset to construct various pose groups for training and testing. With this dataset, we explore the differences in performance between a large variance in viewpoints ranging from frontal facing to extreme side profiles.

4.1.2 Query-Gallery Set Curation

We curate the query and gallery sets based on pose evaluation from our experimental design, ensuring that every identity from the test set is represented. Given the evaluation of a permutation of poses for the query set p_Q and a permutation of poses for the gallery set p_G, we first iterate through each identity in p_Q within the test set and randomly select one image for enrollment into the query set. Next, for the gallery set, we iterate through each identity within the pose group p_G and randomly select 15 images for enrollment into the gallery set. Each query set contains 67 images, representing one image from each identity, while the gallery set contains 1005 images in total, representing 15 images for each of the 67 identities.

For the gallery reduction experiment, we deviate from the previous gallery protocol and aim to reduce the number of images in the gallery set proportionally. Here, the number of images enrolled for an identity is dependent on the gallery poses p_G. We enroll a single image for each unique yaw angle in p_G, where the pitch within the image is randomized and falls between the optimal pitch range. Therefore, the resulting gallery contains N images for N unique yaw angles in p_G.

4.1.3 Model Fine-Tuning

We select a DCNN multi-class classifier ArcFace (Deng et al., 2019) model to represent our benchmark for evaluation. Given ArcFace's high accuracy on unconstrained datasets, we intend to deliver results that represent the current capabilities of DCNN models across the full range of head poses. ArcFace consists of a CNN backbone and an Additive Angular Margin loss function that promotes intra-class compactness and inter-class separation using a margin penalty in a softmax loss function. We attach the ArcFace layer on CNN backbone iResNet-50 (Duta et al., 2020) for fine-tuning and inference to generate face embeddings for a face gallery-matching scenario. Using a pre-trained ArcFace model from the MS1MV3 dataset (Deng et al., 2019) (comprising of frontal poses and achieving ∼95% on the unconstrained dataset IJB-C), we fine-tune the ArcFace model on specific pose groups individually for 10 epochs at a learning rate of $1e^{-1}$. Each face image is downsampled from 512×512 to 112×112 pixels, before being fed to the model. For augmentation experiments, we employ horizontal flips to faces during training to assess performance with mirrored faces.

Table 2: Baseline Accuracy Results From Query Sets Between $-90°$ and 90° Against All Poses Enrolled in Gallery.

Gallery		Query		R-1 Accuracy	Average
Pitch	Yaw	Pitch	Yaw		
-30 to 30	$-90°$ to 90°	-30 to 30	$-90°$ to $-70°$	94.03%	98.51
			$-70°$ to $-45°$	100%	
			$-45°$ to $-15°$	100%	
			$-15°$ to 15°	100%	
			15° to 45°	98.51%	
			45° to 70°	98.51%	
			70° to 90°	98.51%	

4.2 Baseline Performance

As preliminary, we generated baseline results by fixing the gallery set with all available poses for each identity from the M2FPA dataset. By providing all poses, we validate high matching performance for any probe image given that the probe has at least one favorable pose within the set of provided poses. As seen in Table 2, a query set Q_i with a defined set of poses S_Q feeds probe images $q_i \in Q$ to the iResNet-50 CNN backbone to test against the gallery set of all enrolled poses G_{all}, computing the baseline rank-1 metrics. As

Table 3: Rank-1 Accuracy for Query-Gallery Matching Based on Yaw.

Experimental Details	**Query**	**Gallery**							**Average Rank-1**
		$-90°$ to $-70°$	$-70°$ to $-45°$	$-45°$ to $-15°$	$-15°$ to $15°$	$15°$ to $45°$	$45°$ to $70°$	$70°$ to $90°$	
Experiment 1. Initial Pose Sets	$-90°$ to $-70°$	95.52%	67.16%	44.78%	11.94%	16.42%	11.94%	23.88%	38.81%
	$-70°$ to $-45°$	95.52%	100%	98.51%	91.04%	71.64%	73.13%	71.64%	85.93%
	$-45°$ to $-15°$	91.04%	98.51%	98.51%	97.01%	98.51%	89.55%	80.60%	93.39%
	$-15°$ to $-15°$	85.07%	100%	100%	100%	100%	98.51%	91.04%	96.37%
	$15°$ to $45°$	67.16%	95.52%	98.51%	98.51%	100%	98.51%	95.52%	93.39%
	$45°$ to $70°$	68.66%	79.10%	88.06%	91.04%	94.03%	97.01%	98.51%	88.06%
	$70°$ to $90°$	38.81%	22.39%	17.91%	22.39%	53.73%	76.12%	98.51%	47.12%
Experiment 2. Analyzing Between $-45°$ and $45°$	$-45°$ to $45°$	85.07%	97.01%	100%	98.51%	98.51%	100%	94.03%	96.16%
	$-45°$ to $-15° \cup 15°$ to $45°$	89.55%	98.51%	100%	100%	100%	100%	97.01%	97.87%
	$-45°$ to $-22.5° \cup 22.5°$ to $45°$	91.04%	100%	100%	97.01%	98.51%	100%	94.03%	97.23%
	$-45°, 0°, 45°$	95.52%	97.01%	98.51%	100%	100%	100%	92.54%	97.65%

* Pitch is fixed between $-30°$ and $30°$ for all sets in the query and gallery

a result, each query set Q_i scores a relatively high matching accuracy. We see this result as explainable by the presence of similar face poses from the target identity within the gallery, ensuring higher similarity scores in the embedding space.

4.3 Pose Evaluation on Yaw

We compare model performance across the whole yaw range between $-90°$ and $90°$. As a starting point, we build a set of pose groups S for each query and gallery set to assess yaw. We construct each pose set $s_i \in S$ using all pitch angles between $-30°$ and $30°$ and separate yaw within $\sim$20°–30° of the next pose set s_{i+1}.

Given a query set Q_i, which represents the target pose set $s_Q \subset S$ for evaluation, we take each probe image $q_j \in Q_i$ and make a prediction based on the highest cosine similarity with images from the gallery G_j, where the gallery poses are represented by $s_G \subset S$.

As seen in Table 3, we find that when $[-45°, 0°, 45°] \subset s_Q$ results in the highest rank-1 accuracies compared to other angles within the range $[-90°, 90°]$. Pose sets containing frontal angles exhibit consistent accuracy and minimal degradation when extended to extreme yaw angles (up to approximately 90°). Let S represent the initial set of poses, with each pose $s \in S$ corresponding to a yaw angle. We define a restricted subset $S_{\text{frontal}} \subset S$ that contains only the frontal poses within the range $[-45°, 45°]$, where $S_{\text{frontal}} = \{s \in S : -45° \leq s \leq 45°\}$.

Table 4: Pose Ablation on Yaw Between $-45°$ and $45°$.

Poses	Number of Poses	Avg. Rank-1 Accuracy
$-45°$, $-30°$, $-22.5°$, $-15°$, $0°$, $15°$, $22.5°$, $30°$, $45°$	9	96.16%
$-45°$, $-30°$, $-22.5°$, $-15°$, $15°$, $22.5°$, $30°$, $45°$	8	97.87%
$-45°$, $-30°$, $-22.5°$, $22.5°$, $30°$, $45°$	6	97.23%
$-45°$, $0°$, $45°$	**3**	97.65%

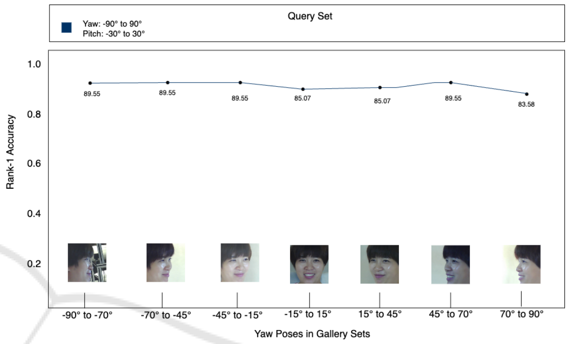


Figure 3: Fine-tuning ArcFace on strictly $-45°$, $0°$, and $45°$ results in uniform recognition performance across yaw.

This subset consistently achieves higher accuracy compared to more extreme angles outside this range. Continuing, we evaluate the pose set $S_{frontal}$ and ablate distinct poses to observe performance stability across G_j. In our last iteration, we constrain our query and training set to a minimal group of poses, $-45°$, $0°$, $45°$. The final pose set containing a frontal and two $\frac{3}{4}$ views produces 97.65% rank-1 accuracy, indicating little to no degradation from the initial 9 poses resulting in 96.16% within $[-45°, 45°]$, as seen in Table 4. In Fig. 3, we visualize the model's stability to pose variation by taking an example identity and verifying high cosine similarity for our suggested poses within the gallery. Regardless of frontal or side profile, performance stabilizes against all yaw angles.

4.4 Pose Evaluation on Pitch

We observe performance stability considering pitch variation. Like yaw, we construct a set of initial poses comprised of negative, positive, and mixed pitch angles.To simplify the evaluation, we constrain the yaw angles to $-45°$, $0°$, and $45°$, based on our prior yaw analysis. For pitch, each probe image $q_i \in Q$ is assigned either 30° (positive pitch) or $-30°$ (negative

Table 5: Rank-1 Accuracy for Query-Gallery Matching Based on Pitch.

Experimental Details	**Query**	**Gallery (columns are separated by yaw)**							**Average Rank-1**
	Yaw: −45°, 0°, 45°	−90° to −70°	−70° to −45°	−45° to −15°	−15° to 15°	15° to 45°	45° to 70°	70° to 90°	
Gallery enrolls all pitches (−30° to 30°)	Pitch: −30°	67.16%	94.03%	97.01%	95.52%	97.01%	94.03%	73.13%	**88.27%**
	Pitch: 30°	55.22%	76.12%	86.57%	83.58%	80.60%	89.55%	53.73%	75.05%
Gallery enrolls only positive pitches (0° to 30°)	Pitch: −30°	71.64%	86.57%	89.55%	89.55%	95.52%	89.55%	68.66%	**84.43%**
	Pitch: 30°	53.73%	79.10%	94.03%	86.57%	94.03%	89.55%	58.21%	79.32%
Gallery enrolls only negative pitches (−30° to 0°)	Pitch: −30°	76.12%	95.52%	98.51%	100%	98.51%	95.52%	73.13%	**91.04**%
	Pitch: 30°	34.33%	55.22%	71.64%	73.13%	67.16%	64.18%	40.30%	57.99%

* Yaw is fixed at −45°, 0°, and 45° for all sets in the query

Table 6: Rank-1 Accuracy for Query-Gallery Matching Using Synthetic Poses.

Experimental Details	**Query**	**Gallery (columns are separated by yaw)**			**Average Rank-1**
	Pitch: −30° to 0°	−45°, 0°, 45° (control)	−45°, 0°, flipped −45°	flipped 45°, 0°, 45°	
Model training enrolls all real images in the training set (No Augmentation)	Yaw: −90° to 90°	91.04%	89.55%	89.55%	90.05%
Model training replaces 45° yaw with flipped −45° view in the training set	Yaw: −90° to 90°	92.54%	91.04%	92.54%	92.04%
Model training replaces −45° yaw with flipped 45° view in the training set	Yaw: −90° to 90°	88.06%	91.04%	89.55%	89.55%

* We run the experiment 5 times and show the fluctuation of results between ∼2-3%

pitch) to compare performance under contrasting conditions during inference. To assess model behavior of varying pitch, we perform gallery matching by cross-testing the query set against galleries with positive, negative, and mixed pitch angles. The results, presented in Table 5, report the rank-1 matching accuracy for probe images with both 30° and −30° pitch.

Probe images with −30° pitch consistently outperform those with 30° across all gallery conditions. Specifically, probe images with −30° pitch achieve an average rank-1 accuracy approximately 13% higher than 30° pitch images when matched against a gallery containing all pitch variations. In addition, −30° pitch probe images exhibit greater resilience to performance degradation when cross-matched with positive pitch images. The accuracy decreases from 91.04% to 84.43%, a modest reduction of ∼7%, when tested against negative versus positive pitch galleries. In contrast, probe images with 30° pitch experience a more substantial drop in performance. The rank-1 accuracy for 30° pitch images drops from an average of 79.32% when matched against negative pitch galleries to 57.99% when cross-matched with positive pitch, representing a decrease of nearly 20%. We identify that probe images with negative pitch demonstrate greater robustness to performance degradation and achieve higher accuracy, regardless of the pitch composition of the gallery. We remark model performance on negative pitch angles complements high performance on the ideal frontal pitch (0°). We infer that poses with 0° or negative pitch are sufficient for DCNN models to capture discriminative features when mapped to the embedding space.

4.5 Evaluation on Synthetic Poses

We exploit the symmetric structure of the human face to leverage horizontal flips of mirrored views. Given a pair of mirrored $\frac{3}{4}$ views (−45° and 45°) and a frontal view (0° yaw), we investigate whether the flipped orientation from a single $|45°|$ yaw angle can effectively replace the opposite view and supplement the complete pair of $\frac{3}{4}$ views. Our hypothesis suggests that this use of synthetic poses through flipping can achieve similar performance to using both authentic −45° and 45° views. Gathering insights from our previous experiments, we use the minimal pitch range (−30° to 0°) and yaw range $(-45°, 0°, 45°)$ to perform augmentation tests. We apply horizontal flipping to replace one of the $\frac{3}{4}$ views and observe the model's performance.

Table 6 presents the results from experiments on galleries augmented with synthetic poses. The query set includes all yaw poses, and the gallery set is constructed using different combinations of real and flipped images. We compare the performance when using flipped images to supplement either the −45° or 45° view, and validate whether there is a significant performance difference between the flipped and non-flipped counterparts. The results show that performance remains uniform across all configurations with only a marginal difference of ∼1-2% between the gallery sets. Our highest performance is achieved when flipping the −45° view to replace 45° in the training set, yielding a mean accuracy of 92.04%. However, the performance difference between this configuration and the others, including the control (using original views), is suggested to be minimal. We attribute the fluctuations in performance (within

Table 7: Image Reduction for Query-Gallery Matching.

Query		Gallery		R-1 Accuracy
Pitch	Yaw	Images/Pose	Yaw	
-30 to 30	-90° to 90°	9	-45° to -45°	80.60%
		6	-45° to $22.5^\circ \cup 22.5^\circ$ to 45°	79.10%
		3	$-45^\circ, 0^\circ, 45^\circ$	82.09%
		2	flipped $45^\circ, 0^\circ, 45^\circ$	80.60%
		2	$-45^\circ, 0^\circ$, flipped -45°	77.61%

[*] Each gallery enrolls one image for each unique yaw angle in an identity's pose set.

a 2-3% range) to the variability during fine-tuning, rather than any preference for flipping a particular yaw direction. We validate these findings by repeating the experiments five times, confirming the consistency of our results. The experiments show that horizontal flips match or slightly improve performance compared to authentic poses. Therefore, we identify the necessary poses for the training and enrollment to be a frontal view (0°) and a single $\frac{3}{4}$ view ($|45^\circ|$).

4.6 Gallery Reduction Evaluation

The final phase of our study reduces the number of images per pose in the gallery set. From previous experiments, we infer that using a yaw set of a single frontal view (0°) and a single $\frac{3}{4}$ view ($|45^\circ|$) with image augmentation is sufficient to produce uniform recognition across all poses. To validate, we devise a query-gallery scenario where each identity enrolls a single image for each unique yaw angle in the identity's pose set. Therefore, given N unique yaw angles, we enroll N images for each identity. During enrollment, the pitch of the face is randomly selected between the proposed optimal pitch range, i.e., -30° to 0°. We evaluate this with a query set randomized with all available poses between yaw -90° to 90° and pitch -30° to 30°.

Table 7 presents our gallery reduction protocol's experimental details and evaluation results. Comparing the performance between pose sets between -45° to 45°, model performance remains relatively constant at $\sim$80% as the poses and number of images reduce. As shown in our results, we validate that the performance is unaffected by image reduction when we reduce the enrollment from 9 poses per ID to only 2 poses per ID. While having more images can be beneficial, there is a lower bound on performance we can expect with minimal images per pose.

5 CONCLUSION

Our main contributions in this study are two-fold. First, we provide a comprehensive analysis of how recognition accuracy varies across pose and augmentation. Furthermore, we implement a training strategy that minimizes the number of facial poses used in training and reduces the need for data collection, using more accessible variations of data. By assessing performance across yaw, pitch, and augmentation, we suggest a training data selection strategy that minimizes poses to yaw angles 0°, a $|45^\circ|$, and a flipped $|45^\circ|$ using horizontal flip augmentation and pitches that are frontal or negative. Following our methodology, we suggest that state-of-the-art DCNN models can result in uniform recognition accuracy across pitch and yaw angles without the necessity of including a large number of poses per subject in the gallery.

REFERENCES

Ahmed, S. B., Ali, S. F., Ahmad, J., Adnan, M., and Fraz, M. M. (2019). On the frontiers of pose invariant face recognition: A review. *Artificial Intelligence Review*, 53(4):2571–2634.

Asthana, A., Marks, T. K., Jones, M. J., Tieu, K. H., and Rohith, M. (2011). Fully automatic pose-invariant face recognition via 3d pose normalization. *2011 International Conference on Computer Vision*.

Baltanas, S.-F., Ruiz-Sarmiento, J.-R., and Gonzalez-Jimenez, J. (2021). Improving the head pose variation problem in face recognition for mobile robots. *Sensors*, 21(2).

Belhumeur, P., Hespanha, J., and Kriegman, D. (1997). Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720.

Bruce, V., Valentine, T., and Baddeley, A. (1987). The basis of the 3/4 view advantage in face recognition. *Applied Cognitive Psychology*, 1(2):109–120.

Chen, G., Shao, Y., Tang, C., Jin, Z., and Zhang, J. (2018). Deep transformation learning for face recognition in the unconstrained scene. *Machine Vision and Applications*, 29(3):513–523.

Cheng, Z., Zhu, X., and Gong, S. (2018). Surveillance face recognition challenge. *arXiv preprint arXiv:1804.09691*.

Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694.

Duta, I. C., Liu, L., Zhu, F., and Shao, L. (2020). Improved residual networks for image and video recognition. *arXiv preprint arXiv:2004.04989*.

Duta, I. C., Liu, L., Zhu, F., and Shao, L. (2021). Improved residual networks for image and video recognition. *2020 25th International Conference on Pattern Recognition (ICPR)*.

Favelle, S. and Palmisano, S. (2018). View specific generalisation effects in face recognition: Front and yaw comparison views are better than pitch. *PLOS ONE*, 13(12).

Gunawan, A. A. and Prasetyo, R. A. (2017). Face recognition performance in facing pose variation. *CommIT (Communication and Information Technology) Journal*, 11(1):1.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

K. Wickrama Arachchilage, S. P. and Izquierdo, E. (2020). Deep-learned faces: A survey. *EURASIP Journal on Image and Video Processing*, 2020(1).

Khaldi, K., Nguyen, V. D., Mantini, P., and Shah, S. (2024). Unsupervised person re-identification in aerial imagery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 260–269.

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M., and Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Li, P., Wu, X., Hu, Y., He, R., and Sun, Z. (2019). M2fpa: A multi-yaw multi-pitch high-quality database and benchmark for facial pose analysis. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Masi, I., Tran, A. T., Hassner, T., Leksut, J. T., and Medioni, G. (2016). Do we really need to collect millions of faces for effective face recognition? *Computer Vision – ECCV 2016*, page 579–596.

Maze, B., Adams, J., Duncan, J. A., Kalka, N., Miller, T., Otto, C., Jain, A. K., Niggel, W. T., Anderson, J., Cheney, J., and et al. (2018). Iarpa janus benchmark - c: Face dataset and protocol. *2018 International Conference on Biometrics (ICB)*.

Müller, M. K., Heinrichs, A., Tewes, A. H., Schäfer, A., and Würtz, R. P. (2007). Similarity rank correlation for face recognition under unenrolled pose. *Advances in Biometrics*, page 67–76.

Nguyen, V. D., Khaldi, K., Nguyen, D., Mantini, P., and Shah, S. (2024). Contrastive viewpoint-aware shape learning for long-term person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1041–1049.

Prince, S. J. and Elder, J. (2006). Tied factor analysis for face recognition across large pose changes. *Procedings of the British Machine Vision Conference 2006*.

Qi, D., Tan, W., Yao, Q., and Liu, J. (2023). Yolo5face: Why reinventing a face detector. *Lecture Notes in Computer Science*, page 228–244.

Rajalakshmi, R. and Jeyakumar, M. K. (2012). A review on classifiers used in face recognition methods under pose and illumination variation. *International Journal of Computer Science Issues (IJCSI)*, 9(6):474–485. Copyright - Copyright International Journal of Computer Science Issues (IJCSI) Nov 2012; Last updated - 2023-11-20.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*.

Torres Pereira, E., Martins Gomes, H., and de Carvalho, J. M. (2014). An approach for multi-pose face detection exploring invariance by training. *Lecture Notes in Computer Science*, page 182–191.

Turk, M. and Pentland, A. (1991). Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yi, D., Lei, Z., Liao, S., and Li, S. (2014). Learning face representation from scratch. *ArXiv*, abs/1411.7923.

Yin, X. and Liu, X. (2018). Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, 27(2):964–975.

Yin, X., Yu, X., Sohn, K., Liu, X., and Chandraker, M. (2019). Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Zhang, X. and Gao, Y. (2009). Face recognition across pose: A review. *Pattern Recognition*, 42(11):2876–2896.