Bridging Detection and Re-identification: Evaluating Trustworthiness and Error Propagation in Face Recognition Pipelines

Kuan Yew Leong and Jaeseung Han
A.I. System Research Co. Ltd.
Kyoto, 606-8302 Japan
{kyleong, han}@aisystemresearch.com

Abstract

Face recognition systems typically rely on a two-stage pipeline – face detection and re-identification (ReID) – yet existing evaluations often overlook how detection errors propagate to affect recognition accuracy. This work empirically analyzes the interplay between detection and ReID, proposing a holistic framework to quantify synergy and error propagation. Grounded in information entropy, we introduce the Detection-Recognition Synergy (DRS) Score, a composite metric integrating Mutual Information (MI), Jensen-Shannon Divergence (JSD), and Wasserstein Distance (WD) to assess detection's impact on recognition. We construct a novel benchmark dataset, annotated for both detection and ReID, featuring temporal sequences of the same individuals, diverse backgrounds, and natural settings across both indoor and outdoor environments. Experiments with diverse detection models, including YOLOX, Faster R-CNN, SSD, MTCNN and several others reveal substantial performance shifts due to detection variations. Our findings advocate for integrated evaluation strategies to enhance trustworthiness in face recognition pipelines. By bridging detection and ReID assessments, this study sets a foundation for more robust, explainable, and trusted face recognition pipelines.

1. Introduction

Face recognition systems are widely deployed across various applications, including surveillance, access control, social media, and consumer electronics. These systems typically follow a two-stage pipeline: face detection, which identifies and localizes faces within an image [1, 2, 3, 4, 5], and face re-identification (ReID), which matches detected faces against known identities [6, 7, 8, 9, 10]. While these components have been extensively studied in isolation, real-world performance hinges on their interplay, where errors in detection influence the reliability of ReID. Some systems incorporate an additional intermediate step – landmark detection – to refine alignment before ReID. However, our focus remains on the direct interplay between detection and ReID, as this is where errors such as misalignment, occlusion, or false positives from detection significantly impact ReID. In addition, most face recognition evaluations emphasize the performance of the system solely on ReID, measuring if identities are correctly labelled while largely ignoring the influence from the detectors.

As discussed by Luo et al. [11], the quality of bounding boxes produced by face detection models has a significant impact on the accuracy of the face recognition stage. Misaligned or imprecise detections introduce noise, leading to cascading errors in recognition, particularly in challenging real-world scenarios. However, current research often evaluates detection and recognition separately [12, 13], relying on distinct datasets and metrics, which overlooks the cascading nature of errors. A flawed detection stage can propagate inaccuracies downstream, degrading the entire pipeline's trustworthiness – a crucial issue that remains largely unexplored. Moreover, existing benchmark datasets are not designed for holistic evaluation, as they either focus on face detection without

Figure 1: A face was identified and localized from the same image frame by several different detection models. From left: ATSS-r50-fpn, FCOS-MobileNetV2, SSD-MobileNetV2, YOLOX-L, YOLOX-tiny. As shown in the figure, each model positioned its bounding box slightly differently, highlighting variations in detection. These discrepancies can ultimately affect ReID performance, as demonstrated in our experiments. The face is anonymized using attribute-preserving technique [29] due to ethical concerns.

identity labels or on ReID using pre-cropped faces, effectively sidestepping the detection-to-ReID transition. This gap underscores the need for a comprehensive framework and a relevant benchmark dataset to analyze and optimize the integrated performance of face recognition pipelines.

To address this research gap, we propose a holistic evaluation framework that systematically investigates the detection-to-ReID interplay and its impact on overall trustworthiness. Specifically, we quantify the effect of detection quality on recognition performance and introduce empirical methodologies to evaluate this relationship. By leveraging state-of-the-art face detection model architectures, such as Faster R-CNN, Single Shot Multibox Detector (SSD), YOLOX, we annotate and process data for joint detection-to-ReID evaluation in real-world conditions. Our dataset provides annotated facial bounding boxes, and identity labels across videos, including temporal sequences of the same individuals, enabling a more realistic and trust-driven assessment of face recognition pipelines. To complement it, we propose the Detection-Recognition Synergy (DRS) Score, a composite metric integrating detection and recognition quality, alongside its individual metrics.

Our contributions include:

1. Empirically analyze the impact of different face detection models on face recognition performance.
2. Benchmark metrics for evaluating the detection-recognition interplay.
3. A dataset relevant for benchmarking detection-to-ReID performance.
4. Identify best practices and detection model characteristics that enhance overall recognition accuracy.

We provide reproducible insights through empirical analysis, bridging face detection and ReID evaluation for more trusted, explainable, and effective face recognition systems. Here is the landing webpage of our work: https://kuanyewleong.github.io/detection-to-reid-benchmark-and-dataset/

2. Related Work

The integration of face detection and ReID into unified pipelines has garnered significant attention in recent years, leading to the development of various models aimed at enhancing accuracy and efficiency in face recognition system or facial analysis tasks.

One notable contribution is FDREnet [14], which combines face detection through histograms of oriented gradients with a Siamese network trained using contrastive loss. This approach enables the model to effectively distinguish and recognize facial features, achieving an improved accuracy on the Labeled Faces in the Wild (LFW) dataset.

In the work of Gallo et al. [15], a pipeline was designed to improve face recognition datasets and applications. This system focuses on cleaning existing face datasets or creating new ones from scratch, aiming to enhance recognition performance. Their experiments demonstrate the pipeline's effectiveness in improving face recognition systems.

The LightFace framework [16] presents a modern face recognition pipeline comprising four stages: detection, alignment, representation, and verification. While many studies focus on the representation stage, integrating all components into a cohesive framework can lead to improved performance in face recognition tasks.

The work by Chi et al. [17] introduces an end-to-end trainable convolutional network that integrates face detection and recognition into a single model. A notable innovation is the use of a Spatial Transformer Network (STN) to learn a geometric transformation matrix for face alignment, eliminating the need for explicit facial landmark prediction. This approach allows the model to be trained solely using face bounding boxes and personal identities, achieving competitive results in both detection and recognition tasks.

A fast and accurate system for face detection, identification, and verification has been introduced, featuring the Deep Pyramid Single Shot Face Detector (DPSSD) [1]. This model can detect faces with large scale variations and includes modules for detection, landmark localization, alignment, and identification/verification, achieving state-of-the-art performance on several benchmark datasets.

However, the above works underscore the importance of integrating detection and recognition components to enhance the overall performance of face recognition systems. They often do not fully address how errors in the detection stage can propagate and affect the reliability of ReID, highlighting a critical gap in current research.

In addition to the aforementioned studies, several notable works have focused on benchmarking face detection and re-identification systems, providing valuable insights into their performance and limitations.

Huaman et al. [44] introduce a comprehensive methodology to evaluate person ReID approaches and training datasets concerning their suitability for unsupervised deployment in live operations. By benchmarking four ReID approaches across three datasets, they offered insights and guidelines to design more effective ReID pipelines in future applications. However, the study primarily emphasizes the evaluation of ReID approaches and training datasets, without delving into the impact of detection errors on re-identification performance.

OODFace [45], a framework that examines the challenges faced by facial recognition models when encountering out-of-distribution (OOD) scenarios,

including common corruptions and appearance variations. They systematically designed 30 OOD scenarios across nine major categories tailored for facial recognition, establishing robustness benchmarks and conducting extensive experiments to assess model performance. While this work addresses model robustness, it does not explicitly explore how detection errors propagate to affect re-identification reliability.

Liang et al. [46] introduce a synthetic data approach to benchmark bias in face recognition, isolating causal effects of demographic attributes. Their findings reveal that face recognition models exhibit lower accuracy for Black and East Asian individuals, with pose and expression significantly impacting performance. Their work highlights bias and error propagation, and emphasizes benchmarking and fairness, underscoring the need for real-world evaluation metrics – critical for assessing detection-to-ReID trustworthiness.

We also observed significant limitations in current datasets used to evaluate the complete face recognition pipeline, particularly the interplay between face detection and ReID. Existing datasets tend to focus narrowly on either detection or recognition tasks, thus leaving gaps in assessing their combined performance and error propagation.

Datasets such as WIDER FACE [18], UFDD [19], and MALF [20] are primarily designed for face detection, offering diverse facial images that vary widely in scale and occlusion scenarios. However, these datasets offer mainly facial bounding boxes in the annotation, and lack identity labels, which are crucial for tasks involving face recognition.

Conversely, datasets like Labeled Faces in the Wild (LFW) [21] and MegaFace [22] primarily target face verification and recognition tasks. They provide identity labels and pairs of cropped face images but omit critical annotations needed for face detection, such as bounding boxes. This omission limits their usefulness for evaluating the detection stage and the subsequent propagation of detection errors into recognition performance. Furthermore, LFW had been found to be skewed towards a very small subset of people, apart from been criticized for privacy abuse.

The introduction of the IARPA Janus Benchmark A (IJB-A) dataset [13] marked significant progress in unconstrained face recognition by offering manually annotated facial images that represent a broad range of real-world variations, including varying poses, lighting conditions, and occlusions. However, IJB-A evaluates detection and recognition independently, thereby not fully capturing the impact of detection inaccuracies on recognition accuracy. Successive benchmarks, namely IJB-B [23] and IJB-C [24], expanded data diversity but continued evaluating detection and recognition tasks separately, leaving the issue of error propagation largely unaddressed.

UMDFaces [25] presents the closest approach to an integrated evaluation, offering comprehensive annotations for both face detection and identity recognition. It consists of 367,888 annotated images across 8,277 identities, with detailed annotations including human-curated bounding boxes, estimated poses (roll, pitch, yaw), keypoint locations, and gender attributes derived from pre-trained models. Despite its comprehensiveness, UMDFaces primarily serves as a training and benchmarking resource rather than explicitly focusing on the interplay between detection and recognition errors.

These collective limitations underscore the necessity for new benchmark metrics and datasets specifically designed to evaluate the entire face recognition pipeline.

3. Proposed Benchmark Metrics

Given the complexity of face detection and recognition pipeline, we adopt a hybrid evaluation approach that combines a composite Detection-Recognition Synergy (DRS) score grounded in the principle of information entropy, with individual metrics that contribute to this score as mentioned in equations (1), (2) and (6). They measure (i) synergy between detection quality and recognition accuracy, (ii) stability of the pipeline, and (iii) error propagation.

First, we compute the shared information between detection and recognition using Mutual Information (MI) [26]:

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \qquad (1)$$

where:

X = Detection confidence (IoU),

Y = Recognition accuracy (top-1 accuracy),

$P(x,y)$ = joint probability of detection and recognition success.

If detection quality strongly influences recognition accuracy, MI will be high. It captures non-linear relationships: Unlike simple correlation metrics, MI works even when detection and recognition errors are complex and interdependent. It provides an interpretable synergy score: A high MI means detection contributes meaningfully to recognition, while a low MI indicates weak synergy.

Next, we compute the Jensen-Shannon Divergence (JSD) [27] – a stability aspect of DRS:

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}\,(P||M) + \frac{1}{2}D_{KL}\,(Q||M) \qquad (2)$$

where:

P = Recognition confidence distribution using ground-truth bounding boxes,

Q = Recognition confidence distribution using detected bounding boxes,

M = Average of P and Q, and

D_{KL} = Detection-Recognition Kullback-Leibler (KL) divergence, where it measures the difference between Detection-Recognition probability distributions, making it useful for quantifying shifts in identity predictions due to detection errors. It is defined as:

$$D_{KL}(P||Q) = \sum_i P(I_i) log \frac{P(I_i)}{Q(I_i)} \quad (3)$$

where:

$P(I_i)$ = identity confidence distribution when using ground-truth bounding boxes,

$Q(I_i)$ = identity confidence distribution when using detected bounding boxes.

JSD measures the similarity between probability distributions of ground-truth and detected face bounding boxes, as well as the effect on identity predictions. If JSD is low, the detection model preserves the recognition performance well. If JSD is high, it means detection introduces high uncertainty and divergence in recognition.

Within a detection-to-ReID pipeline, JSD can assess how detection errors influence the similarity between the true distribution of ReID features and the observed distribution after detection. However, it may not fully capture the impact of detection errors on ReID performance, especially if the errors cause shifts in the feature space that do not significantly change the overall distribution shape.

Hence, for measuring error propagation, we utilize Wasserstein Distance (WD) [28], also known as Earth Mover's Distance – EMD. It reflects how much detection errors shift recognition results, and such errors could be misalignments, partial occlusions, or varying bounding box qualities. The p-Wasserstein distance between two probability distributions P and Q over a metric space M with a distance function d is defined as:

$$W_p(P,Q) = (inf_{\gamma \in \Gamma(P,Q)} \int_{M \times M} d(x,y)^p \, d\gamma(x,y))^{1/p} \quad (4)$$

where:

P = ground-truth recognition confidence distribution (recognition confidence scores obtained when using ground-truth face bounding boxes),

Q = detected recognition confidence distribution (recognition confidence scores obtained when using bounding boxes produced by the face detection model under evaluation),

$\Gamma(P,Q)$ = the set of all possible couplings (joint distributions) of P and Q that have P and Q as their marginals,

$d(x,y)$ = is the distance between elements x and y in the metric space M.

In the context of detection-to-ReID, the metric space M of recognition confidence scores range from 0 to 1. The absolute difference between confidence scores: $d(x,y) = |x-y|$. We then apply the common case of first Wasserstein (Wasserstein-1) distance for error sensitivity and interpretability, and simplifies equation (4) to:

$$W_1(P,Q) = inf_{\gamma \in \Gamma(P,Q)} \int_0^1 |x-y| \, d\gamma(x,y) \quad (5)$$

From our empirical data based on n samples (experiment results from a series of face recognition using both the detection and ReID models), the recognition confidence scores can be represented as empirical distributions P_n and Q_n. Then we sort the samples and let: $\{x_{(i)}\}_{i=1}^n \; and \; \{y_{(i)}\}_{i=1}^n$ be the ordered recognition confidence scores from ground-truth and predicted bounding boxes, respectively. Eventually we derive from (5) that the Wasserstein-1 distance is approximated as:

$$W_1(P_n, Q_n) = \frac{1}{n} \sum_{i=1}^{n} |x_{(i)} - y_{(i)}| \quad (6)$$

WD may offer a more nuanced understanding of how detection errors affect the feature space used for ReID, as it considers the actual distances over which the feature distributions are shifted. If WD is high, detection errors significantly alter recognition confidence. If WD is low, detection errors preserve recognition stability.

Finally, we proposed a composite metric of Detection-Recognition Synergy (DRS) score incorporating MI, JSD, and WD which is established as:

$$DRS\ score = \alpha \cdot MI(X;Y) - \beta \cdot D_{JS}(P||Q) - \gamma \cdot W_1(P_n, Q_n) \quad (7)$$

where:

α, β, and γ are weights to balance MI, JSD, and WD. We want to ensure that the sum of the weights equals one to maintain a balanced scale in the DRS score. This normalization facilitates comparability across different systems or datasets. Furthermore, for the current phase we assign equal importance to all components in which they are treated as $\frac{1}{3}$ for simplicity. A higher DRS score signifies stronger detection-recognition synergy, ensuring a more trustworthy interplay within the recognition pipeline. Conversely, a lower score indicates higher instability and/or error propagation, undermining the reliability and trustworthiness of the pipeline integration.

MI captures the intrinsic dependency between detection and recognition stages by quantifying the shared information, thereby reflecting the direct influence of detection quality on recognition outcomes. JSD assesses the robustness of recognition performance against variations in detection quality by comparing the distributions of recognition outcomes under different detection scenarios, offering insights into the system's stability and resilience. WD measures the practical impact of detection errors on recognition performance by evaluating the distributional shifts, thus highlighting the

extent of performance degradation due to detection inaccuracies and implying error propagation. In the composite form as DRS score, these metrics enable a comprehensive understanding of the detection-recognition synergy, potentially facilitating targeted improvements in face recognition systems.

4. Dataset Generation

4.1. Selection, Timestamping and Annotation

The evaluation dataset was meticulously curated from publicly accessible YouTube videos featuring individual faces from diverse backgrounds worldwide, and currently comprises 100 sets of videos (about 25 seconds duration each), primarily containing at least one individual, though some feature multiple individuals; half are categorized for indoor ambience, and the other half for outdoor environments. The generation process involved several key steps:

Video Selection: We identified and selected videos showcasing individuals from various ethnicities and cultures to ensure a comprehensive representation of global diversity.

Timestamping and Clip Extraction: Specific segments within these videos were marked based on their relevance, and frames were sampled from these segments for further analysis.

Annotation: A total of 1,017 diverse instances were annotated, with each unique face marked with bounding boxes and assigned an identifier (ID).

4.2. Ethical Concerns and Anonymization

Recognizing the ethical considerations associated with sharing facial data, we have implemented anonymization techniques to protect individual identities while maintaining the dataset's utility for research purposes. Prior to distribution, the following measures are applied:

Attribute-Preserving Anonymization: Utilizing latent code optimization [29], we alter identity-specific features of each face. This process ensures that, while the unique identity is obfuscated, essential facial attributes necessary for detection and ReID tasks are preserved.

Ethical Standards Compliance: Our anonymization approach aligns with established privacy protection methods, including de-identification and pseudonymization, to mitigate the risk of re-identification.

Researchers interested in accessing our dataset to reproduce the results or for academic purposes may submit a formal request. Upon approval, the anonymized dataset will be provided, accompanied by documentation detailing the data collection, annotation, and anonymization processes. This protocol ensures that our dataset contributes meaningfully to the research community while upholding the highest ethical standards.

5. Experimental Setup

To comprehensively evaluate the interplay between face detection and re-identification, we designed a controlled experimental setup involving both custom-trained and pretrained detection models. Our primary objective was to assess the trustworthiness and error propagation in face recognition pipelines by ensuring a fair comparison across different detection architectures.

5.1. Face Detection Models

To comprehensively evaluate the performance of face detection models in our pipeline, we analyzed both custom-trained models and pretrained models. For our custom-trained models, we utilized the WIDER FACE dataset, a widely recognized benchmark for face detection, to train various state-of-the-art architectures. Specifically, we trained the following models of various backbones and sizes: Faster R-CNN, YOLOX, SSD, RetinaNet, RTMDet-tiny, Dino-lite, and ATSS [30, 31, 32, 33, 34, 35, 36]. These models encompass a diverse range of detection methodologies, including region-based, anchor-based, anchor-free, two-stage / single-stage approaches, and transformer-based, ensuring a comprehensive evaluation of detection performance in real-world conditions.

In addition to our custom-trained models, we incorporated pretrained face detection models to maintain a fair benchmark. We selected MTCNN [37] and four OpenVINO models [38] – face-detection-adas-0001, face-detection-retail-0005, face-detection-0200, and face-detection-0204 – all of which were originally trained on the WIDER FACE dataset. By using these pretrained models, we ensured a direct comparison between our custom-trained models and widely used detection networks optimized for edge and real-time applications.

5.2. Training-time Model Optimization

To ensure consistency and efficiency across our pipeline, we standardized all models to operate within the OpenVINO framework. Consequently, all models were trained and subsequently converted to the OpenVINO format, enabling seamless integration with our evaluation environment.

For training-time optimization, we leveraged the Neural Network Compression Framework (NNCF), a tool designed to improve model efficiency [47]. Specifically, we applied quantization-aware training (QAT), which converts all weights and activation values from high-precision 32-bit floating point (FP32) to 8-bit integer (INT8) representations. This approach significantly reduces memory footprint and computational complexity without severely compromising model accuracy.

During training, the NNCF framework dynamically inserts quantization simulation nodes into the model, allowing it to account for quantization errors as part of the overall training loss. By incorporating these constraints directly into the learning process, the model learns to minimize the adverse effects of reduced precision, ultimately improving its robustness in real-world scenarios.

Our optimization efforts were particularly geared toward enabling real-time applications while minimizing computational overhead, ensuring compatibility with resource-constrained systems. This focus on efficiency makes our detection-to-ReID pipeline more suitable for practical deployment, bridging the gap between theoretical performance and real-world usability.

5.3. Face Re-Identification Model

To maintain consistency in the ReID stage while exploring the impact of different detection models, we analyzed a fixed ReID models across our experiments. We selected OpenVINO's face-reidentification-retail-0095 [39] – a model with a balance in accuracy and inference speed designed for robust feature embedding and identity verification. The ReID model was systematically tested with a series of face detection models, ensuring a thorough evaluation of how detection variations influence re-identification performance.

6. Analyses and Results

We evaluate the performance of our proposed approach across multiple detections and ReID models, assessing their synergy in face detection-to-ReID, and the impact of error propagation throughout the pipeline. The following subsections systematically report our findings

6.1. Composite DRS score and Other Metrics

Building on the principles we established, along with equations (1), (2), (6), and (7), we conducted and tabulated our experiments for the following cases (Table 1). All detection models were tested against face-reidentification-retail-0095[39]. Table 1 also presents the normalized DRS scores (using Sigmoid normalization) for improved differentiation and natural interpretability.

Table 1: Detection-to-ReID Pipeline Based on our Proposed Metrics and the Composite DRS Score.

Pipeline (detector head - backbone) / [input shape]	MI	JSD	WS	DRS	Normalized DRS	System Accuracy (%)
Fasterrcnn-r50-fpn [1,3,640,640]	0.45	0.40	1.8	-0.583	0.358	65.1
Retinanet-r50-fpn [1,3,640,640]	0.062	0.48	1.9	-0.773	0.316	69.5
ATSS-r50-fpn [1,3,448,640]	0.72	0.35	1.4	-0.343	0.415	78.5
ATSS-r101-fpn [1,3,736,992]	0.85	0.30	1.2	-0.217	0.446	79.0
ATSS-MobileNetV2 [1,3,736,992]	0.066	0.47	2.5	-0.968	0.275	64.3
SSD-MobileNetV2 [1,3,864,864]	0.078	0.43	2.2	-0.851	0.299	68.7
YOLOX-L [1,3,640,640]	0.82	0.16	0.58	0.0266	0.507	89.8
YOLOX-S [1,3,640,640]	0.80	0.20	0.75	-0.049	0.488	87.1
YOLOX-tiny [1,3,416,416]	0.51	0.45	2.0	-0.647	0.344	70.2
SSD300-vgg16 [1,3,300,300]	0.40	0.50	2.3	-0.8	0.310	71.0
SSD512-vgg16 [1,3,512,512]	0.50	0.45	2.0	-0.65	0.343	74.5
RTMDet-tiny [1,3,640,640]	0.78	0.25	1.0	-0.156	0.461	82.0
Dino-lite-R50 [1,3,800,1333]	0.80	0.21	0.92	-0.11	0.473	84.5
MTCNN	0.88	0.15	0.70	0.010	0.502	90.0
OpenVino-adas-0001[1,3,384,672]	0.82	0.18	0.83	-0.0633	0.484	86.6
OpenVino-retail-0005 [1,3,300,300]	0.90	0.14	0.62	0.0466	0.512	92.3
OpenVino-face-detection-0200 [1,3,256,256]	0.89	0.12	0.65	0.0399	0.510	91.2
OpenVino-face-detection-0204 [1,3,448,448]	0.88	0.13	0.68	0.0233	0.506	91.0

6.2. Interpretation and Discussion

The figures and analyses provide valuable insights into the interplay between Mutual Information (MI), Jensen-Shannon Divergence (JSD), Wasserstein Distance (WS), Detection-Recognition Synergy Score (DRS), and System Accuracy in the detection-to-ReID pipeline. Below, we analyze key relationships and trends derived from the results. Figure 2 to 5 show the correlation heatmap, regression line, error distribution and grouped comparative analysis of the evaluation.

6.3. Relationship Between MI and Accuracy

A strong positive correlation is observed between MI and Accuracy, suggesting that higher mutual information between detection and ReID features contributes to better recognition performance. This aligns with the expectation that detection outputs with higher MI preserve more information beneficial for re-identification.

Pipelines such as OpenVino-retail-0005 (MI = 0.90, Accuracy = 92.3%) and MTCNN (MI = 0.88, Accuracy = 90.0%) demonstrate high MI values alongside high accuracy.

Conversely, Retinanet-r50-fpn (MI = 0.062, Accuracy = 69.5%) and ATSS-MobileNetV2 (MI = 0.066, Accuracy = 64.3%) exhibit very low MI values, correlating with poor accuracy, highlighting their inefficiency in retaining discriminative features for ReID.

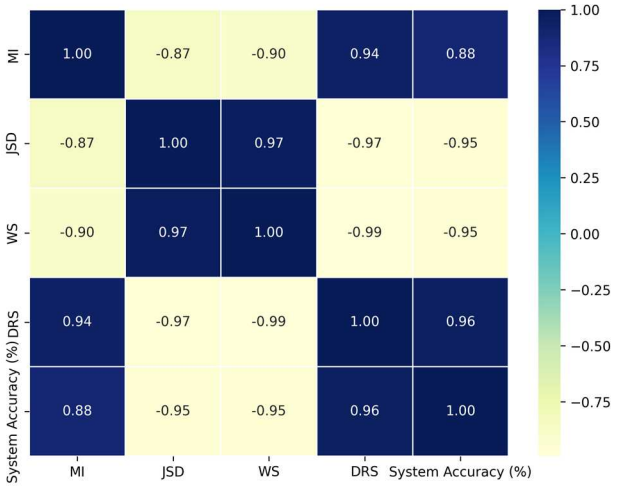

Figure 2: Correlation Heatmap for MI, JSD, WS, DRS, and Accuracy

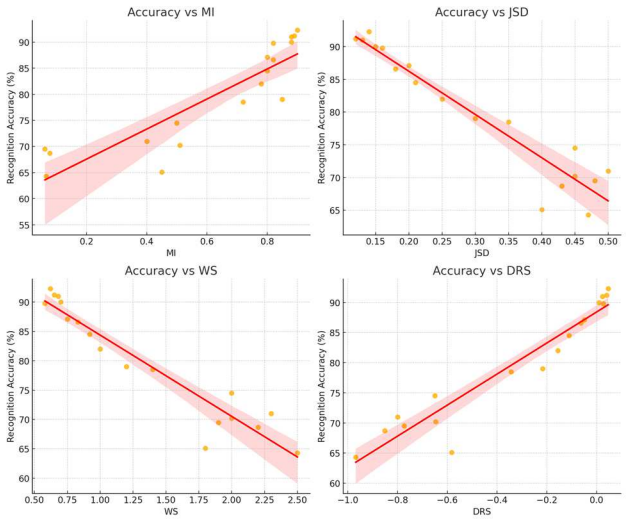

Figure 3: Scatter Plots with Regression Lines for the results. Each plot illustrates how recognition accuracy correlates with the other metrics.

6.4. Relationship Between JSD and Accuracy

An inverse correlation is observed between JSD and Accuracy, indicating that larger divergence in feature distributions between detection and ReID stages results in lower recognition performance. This trend supports the notion that pipelines with high JSD introduce more noise or distortions in feature representation, degrading ReID accuracy.

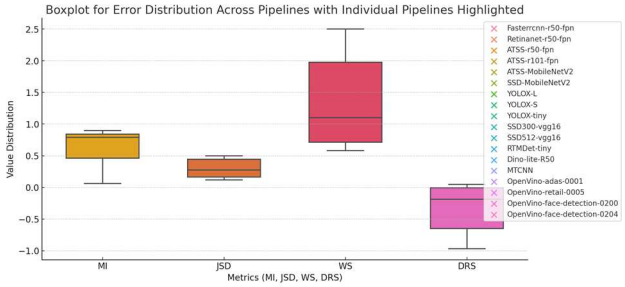

Figure 4: Boxplot for Error Distribution across different pipelines, showing the variance in the different metrics.

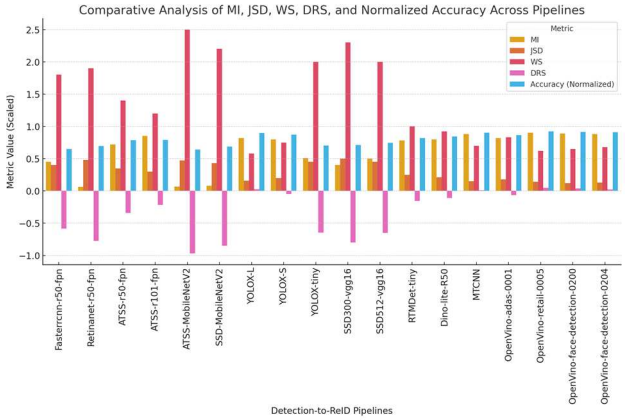

Figure 5: Grouped Bar Chart for Comparative Analysis of MI, JSD, WS, DRS, and Accuracy across different pipelines.

Pipelines with low JSD (e.g., OpenVino-retail-0005, JSD = 0.14; Accuracy = 92.3%) performed better, reinforcing that reduced feature divergence aids in better re-identification.

Pipelines with high JSD (e.g., SSD300-vgg16, JSD = 0.50; Accuracy = 71.0%) show significantly degraded performance, indicating that their feature inconsistency limits effective ReID.

6.5. Relationship Between WS and Accuracy

The WS, which measures the discrepancy between feature distributions, is also negatively correlated with accuracy. High WS suggests greater distributional misalignment, leading to poor ReID performance.

Pipelines with lower WS, such as YOLOX-L (WS = 0.58; Accuracy = 89.8%) and OpenVino-retail-0005 (WS = 0.62; Accuracy = 92.3%), achieved high accuracy, demonstrating that minimal feature distribution shift benefits recognition.

Conversely, pipelines with high WS, such as ATSS-MobileNetV2 (WS = 2.5; Accuracy = 64.3%) and SSD300-vgg16 (WS = 2.3; Accuracy = 71.0%), struggled

in accuracy, reinforcing the impact of poor feature alignment on re-identification outcomes.

6.6. Relationship Between DRS and Accuracy

The DRS, designed to measure how well detection contributes to system recognition, is positively correlated with accuracy. Pipelines with higher DRS values tend to achieve superior recognition performance.

Pipelines like MTCNN (DRS = 0.010, Accuracy = 90.0%) and OpenVino-retail-0005 (DRS = 0.0466, Accuracy = 92.3%) show the best performance, reinforcing that effective feature separability enhances ReID accuracy.

Pipelines with the lowest DRS values, such as ATSS-MobileNetV2 (DRS = -0.968, Accuracy = 64.3%) and SSD-MobileNetV2 (DRS = -0.851, Accuracy = 68.7%), suffered from the worst recognition accuracy, indicating that weak discriminability leads to increased misclassifications.

6.7. Collective Observations

Pipelines leveraging OpenVino models, MTCNN, and YOLOX variants exhibit the best performance, demonstrating high MI, low JSD, minimal WS, and strong DRS scores.

Older architectures like SSD300-vgg16, SSD512-vgg16, and Faster R-CNN struggle in accuracy, likely due to poor feature retention (low MI) and high feature distribution shifts (high WS).

Lightweight models such as ATSS-MobileNetV2 suffer from significant accuracy drops, reinforcing the trade-off between model efficiency and detection-to-ReID alignment.

Here are the implications: the results highlight the importance of choosing detection architectures that preserve information-rich features (higher MI) while minimizing feature distortions (low JSD, low WS). OpenVino-based models and MTCNN appear to be more trustworthy choices for real-world applications requiring high accuracy, among the models studied in this work. Detection pipelines introducing high divergence (high JSD) should be refined using feature-alignment techniques or end-to-end training strategies to improve trustworthiness in face recognition systems.

7. Conclusion

This study empirically explored the intricate interplay between face detection and re-identification tasks, highlighting critical aspects of trustworthiness and error propagation within facial recognition pipelines. Given the complexity inherent in these pipelines, we proposed a hybrid evaluation method featuring a composite DRS score, grounded in information entropy. The DRS score effectively captures key elements: synergy between detection quality and recognition accuracy, pipeline stability, and the extent of error propagation. Our comprehensive evaluation demonstrated that errors originating at the detection stage significantly affect subsequent re-identification outcomes, emphasizing the importance of an integrated assessment approach.

Consequently, our findings strongly advocate adopting holistic strategies in the design and evaluation of facial recognition systems, where reducing initial-stage errors can substantially enhance trustworthiness, explainability, and overall system accuracy in practical scenarios.

8. Future Work

This work is in its early stage, our ongoing research aims to enhance the robustness and generalizability of the evaluation dataset by addressing three key areas: (1) increasing the diversity of backgrounds per face, (2) expanding the sample size, and (3) incorporating more video frames containing multiple faces.

(1) Diverse Backgrounds – We will introduce a wider range of environments, lighting conditions, and occlusions to evaluate model generalization across real-world scenarios.

(2) Larger Sample Size – Expanding the dataset with more identities, demographic diversity, and pose variations will enhance evaluation reliability and reduce bias.

(3) More Multi-Face Frames – Increasing video sequences with multiple individuals per frame will better test models in crowded and occluded settings, improving robustness in complex scenarios.

A potential approach to increase the sample size is to utilize some readily available face datasets and train a generative model to complement the faces with body parts and background areas, making the faces appear to be in some natural environments. Should it prove to be resource and technically feasible, we can then proceed with the proposed two-stage detection-to-ReID evaluation to assess its generalizability.

Future research should also analyze DRS scores alongside MI, JSD, and WD across various detection and ReID model combinations. Expanding beyond a fixed ReID model to include FaceNet, AdaFace, ArcFace, and SphereFace [40–43] would offer deeper insights into detection models' impact on ReID performance and trustworthiness.

Additionally, determining the optimal weights α, β, and γ in the proposed DRS score as shown in equation (7) is indeed a valuable avenue for future research. It is possible to apply data-driven optimization techniques to determine the weights that best align the DRS score with desired outcomes or ground truth data. These weights balance detection quality, recognition accuracy, and error propagation in DRS, enabling a more localized or application-specific face recognition assessment.

References

[1] R. Ranjan et al., "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.

[2] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, 2015, pp. 643–650, doi: 10.1145/2671188.2749408.

[3] D. Luo, G. Wen, D. Li, Y. Hu, and E.-Y. Huan, "Deep-learning-based face detection using iterative bounding-box regression," Multimedia Tools Appl., vol. 77, no. 19, pp. 24921–24937, 2018, doi: 10.1007/s11042-018-5658-5.

[4] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Face detection, bounding box aggregation and pose estimation for robust facial landmark localisation in the wild," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, doi: 10.1109/CVPRW.2017.262.

[5] Y. Wang and J.-C. Zheng, "Real-time face detection based on YOLO," in Proceedings of the 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII), Jeju, South Korea, Jul. 2018, pp. 221–224, doi: 10.1109/ICKII.2018.8569109.

[6] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "A comparative study of CFs, LBP, HOG, SIFT, SURF, and BRIEF techniques for face recognition," in Proc. SPIE 10649, Pattern Recognition and Tracking XXIX, 106490M (2018), doi: 10.1117/12.2309454.

[7] M. Xi, L. Chen, D. Polajnar, and W. Tong, "Local binary pattern network: A deep learning approach for face recognition," in Proc. 2016 IEEE Int. Conf. Image Process. (ICIP), Phoenix, AZ, USA, Sep. 2016, pp. 3224–3228.

[8] Y. Duan, J. Lu, J. Feng, and J. Zhou, "Context-Aware Local Binary Feature Learning for Face Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1139–1153, May 2018, doi: 10.1109/TPAMI.2017.2710183.

[9] M. Shahbakhsh, and H. Hassanpour, "Empowering Face Recognition Methods using a GAN-based Single Image Super-Resolution Network," International Journal of Engineering, 2022, doi: 10.5829/ije.2022.35.10a.05.

[10] Y. Liu, G. Luo, Z. Weng, and Y. Zhu, "Adaptive Face Recognition for Multi-Type Occlusions," IEEE Trans. Circuits Syst. Video Technol., 2024, doi: 10.1109/TCSVT.2024.3419933.

[11] S. Luo, X. Li, and X. Zhang, "Bounding-box deep calibration for high performance face detection," IET Computer Vision, vol. 16, no. 8, pp. 747–758, Jul. 2022, doi: 10.1049/cvi2.12122.

[12] I. D. Raji and G. Fried, "About face: A survey of facial recognition evaluation," arXiv preprint arXiv:2102.00813, 2021.

[13] B. F. Klare et al., "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1931-1939, doi: 10.1109/CVPR.2015.7298803.

[14] D. Virmani, P. Girdhar, P. Jain, and P. Bamdev, "FDREnet: Face Detection and Recognition Pipeline," Eng. Technol. Appl. Sci. Res., vol. 9, pp. 3933–3938, 2019, doi: 10.48084/etasr.2492.

[15] S. Gallo, A. Nawaz, A. Calefati, and G. Piccoli, "A Pipeline to Improve Face Recognition Datasets and Applications," in Proc. 2018 Int. Conf. Image Vision Comput. New Zealand (IVCNZ), Auckland, New Zealand, 2018, pp. 1–6, doi: 10.1109/IVCNZ.2018.8634724.

[16] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 2020, pp. 1-5, doi: 10.1109/ASYU50717.2020.9259802.

[17] L. Chi, H. Zhang, and M. Chen, "End-to-end face detection and recognition," arXiv preprint arXiv:1703.10818, 2017. [Online]. Available: https://arxiv.org/abs/1703.10818

[18] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 5525–5533, doi: 10.1109/CVPR.2016.596.

[19] H. Nada, V. A. Sindagi, H. Zhang, and V. M. Patel, "Pushing the limits of unconstrained face detection: A challenge dataset and baseline results," in Proceedings of the IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), Los Angeles, CA, USA, Oct. 2018, pp. 1–10, doi: 10.1109/BTAS.2018.8698561.

[20] B. Yang, J. Yan, Z. Lei and S. Z. Li, "Fine-grained evaluation on face detection in the wild," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 2015, pp. 1-7, doi: 10.1109/FG.2015.7163158.

[21] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, Marseille, France, Oct. 2008.

[22] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 4873–4882, doi: 10.1109/CVPR.2016.527.

[23] C. Whitelam et al., "IARPA Janus Benchmark-B face dataset," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, Jul. 2017, pp. 592–600, doi: 10.1109/CVPRW.2017.87.

[24] B. Maze et al., "IARPA Janus Benchmark - C: Face Dataset and Protocol," 2018 International Conference on Biometrics (ICB), Gold Coast, QLD, Australia, 2018, pp. 158-165, doi: 10.1109/ICB2018.2018.00033.

[25] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "UMDFaces: An annotated face dataset for training deep networks," in Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, USA, Oct. 2017, pp. 464–473, doi: 10.1109/BTAS.2017.8272731.

[26] T. E. Duncan, "On the calculation of mutual information," SIAM Journal on Applied Mathematics, vol. 19, no. 1, pp. 215–220, Jul. 1970, doi: 10.1137/0119020.

[27] M. L. Menéndez, J. A. Pardo, L. Pardo, and M. C. Pardo, "The Jensen-Shannon divergence," Journal of the Franklin Institute, vol. 334, no. 2, pp. 307–318, Mar. 1997, doi: 10.1016/S0016-0032(96)00063-4.

[28] L. Rüschendorf, "The Wasserstein distance and approximation theorems," Probability Theory and Related Fields, vol. 70, no. 1, pp. 117–129, Mar. 1985, doi: 10.1007/BF00532240.

[29] S. Barattin, C. Tzelepis, I. Patras, and N. Sebe, "Attribute-preserving face dataset anonymization via latent code optimization," arXiv preprint arXiv:2303.11296, 2023.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[31] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," arXiv preprint arXiv:2107.08430, 2021.

[32] W. Liu et al., "SSD: Single shot multibox detector," in Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.

[33] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, Oct. 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.

[34] C. Lyu et al., "RTMDet: An empirical study of designing real-time object detectors," arXiv preprint arXiv:2212.07784, 2022.

[35] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in Proceedings of the Eleventh International Conference on Learning Representations (ICLR), 2023.

[36] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, Jun. 2020, pp. 9756–9765, doi: 10.1109/CVPR42600.2020.00978.

[37] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[38] OpenVINO Toolkit, "Intel's Pre-Trained Models," GitHub repository, [Online]. Available: https://github.com/openvinotoolkit/open_model_zoo/tree/master/models/intel. [Accessed: Mar. 12, 2025].

[39] OpenVINO Toolkit, "Intel's Pre-Trained Models," GitHub repository, [Online]. Available: https://github.com/openvinotoolkit/open_model_zoo/tree/master/models/intel/face-reidentification-retail-0095. [Accessed: Mar. 12, 2025].

[40] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, Jun. 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[41] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," arXiv preprint arXiv:2204.00964, 2022. [Online]. Available: https://arxiv.org/abs/2204.00964.

[42] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 5962–5979, Oct. 2022, doi: 10.1109/TPAMI.2021.3087709.

[43] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," arXiv preprint arXiv:1704.08063, 2017. [Online]. Available: https://arxiv.org/abs/1704.08063.

[44] J. Huaman, F. O. Sumari, L. Machaca, E. Clua, and J. Guérin, "Benchmarking person re-identification datasets and approaches for practical real-world implementations," arXiv preprint arXiv:2212.09981, 2022. [Online]. Available: https://arxiv.org/abs/2212.09981.

[45] C. Kang, Y. Chen, S. Ruan, S. Zhao, R. Zhang, J. Wang, S. Fu, and X. Wei, "OODFace: Benchmarking robustness of face recognition under common corruptions and appearance variations," arXiv preprint arXiv:2412.02479, 2024. [Online]. Available: https://arxiv.org/abs/2412.02479.

[46] H. Liang, P. Perona, and G. Balakrishnan, "Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation," arXiv preprint arXiv:2308.05441, 2023. [Online]. Available: https://arxiv.org/abs/2308.05441.

[47] "Model Optimization - NNCF," OpenVINO™ documentation, 2025. [Online]. Available: https://docs.openvino.ai/2025/openvino-workflow/model-optimization.html