# Quality-Aware Person Re-Identification for Improved Fusion of Visual and Soft-Biometric Cues

Lina El Arajna, Nino Lalanne-Tisné and Donatien Madjibaye*

Polytech Nantes, Nantes Université

**Abstract**

*Person re-identification (ReID) systems based only on visual cues can degrade in realistic conditions (low resolution, blur, occlusions, illumination changes), where identity evidence becomes ambiguous [1]. This project builds on the MSPA architecture presented in [2] by implementing the visual branch (APN) and the soft-attribute branch (APAN) to leverage complementary soft biometrics for more robust ReID [3]. In addition, we propose to modify the fusion strategy by introducing a third branch dedicated to Image Quality Assessment (IQA) proposed by [4] through an* identifiability *score. Unlike human-oriented quality metrics, identifiability is an unsupervised, model-centric measure that estimates how reliable a query image is for ReID: we synthetically perturb the query (e.g., blur, brightness changes, occlusions, noise, background variations), extract deep features from a pretrained ResNet-50 [5, 6] for the original and perturbed versions, and compute a consistency score between the resulting feature vectors. The identifiability score is fed into a lightweight MLP to produce a fusion weight, which is used to weight the two branches' contributions. We evaluate the approach on Market-1501 [7], a popular ReID dataset, using standard ReID metrics (mAP, Rank-1 and Rank-5).*

## I Introduction

Person re-identification (ReID) is defined as the task of retrieving and matching images of a specific individual across a distributed network of non-overlapping cameras [1]. As a cornerstone of modern multi-camera video surveillance and forensic analysis, ReID aims to maintain identity continuity in complex environments where traditional biometric modalities, such as face or iris recognition, are often unavailable due to low image resolution or distance from the sensor [8].

The fundamental principle of ReID lies in finding a robust and discriminative representation of a person's appearance that remains invariant to significant intra-class variations. Historically, the field has evolved from handcrafted descriptors (focusing on color histograms and texture patterns) to deep learning architectures [1]. Modern approaches typically rely on Convolutional Neural Networks (CNNs) or Vision Transformers to learn high-level semantic features [1]. These models generally follow two main paradigms: global feature learning, which captures the overall silhouette and color distribution, and local/part-based learning, which focuses on specific body regions to handle partial occlusions and pose misalignments [1, 9].

Despite achieving impressive performance on curated benchmarks, the reliability of ReID systems often degrades in real-world deployments. This performance gap is primarily caused by several "nuisance factors" inherent to unconstrained surveillance settings: extreme viewpoint changes, low spatial resolution, background clutter, and severe illumination shifts. Furthermore, different individuals may wear similar clothing (inter-class similarity), while the same person's appearance can change drastically across different cameras (intra-class variation), leading to visually ambiguous matches. It is important to note that while some recent research addresses the "clothes-changing" ReID problem where individuals change outfits over long periods [10], this work focuses on the standard Non-Clothes-Changing (NCC) setting, where short-term visual consistency is assumed. In such NCC scenarios, pixel-level information alone is often still insufficient for confident identification due to the aforementioned environmental noise [1].

To mitigate these limitations, a promising research direction is the integration of *soft biometric* attributes as discussed in [1, 3]. Unlike hard biometrics, soft biometrics provide high-level semantic cues that are easily understandable by humans and less sensitive to fine-grained pixel noise. These include categorical labels such as gender and age group, as well as physical characteristics like hair length, sleeve length, clothing type, and the presence of accessories (e.g., backpacks, hats, or handbags), as illustrated in Fig 1 where a single individual is captured from multiple viewpoints along with their corresponding semantic descriptors. By complementing visual features with these semantic descriptors, models can effectively prune the search space and disambiguate candidates that share similar global appearances but differ in key details [1, 3, 11].

In this work, we build upon the Multi-Scale Pyra-

*Supervisor: Rebiha Souadih (*rebiha.souadih@univ-nantes.fr*).

| | | | |
|---|---|---|---|
| **Gender** | Female | **Carrying backpack** | No |
| **Hair length** | Short hair | **Carrying bag** | Yes |
| **Sleeve length** | Short sleeve | **Carrying handbag** | No |
| **Lower-body length** | Short | **Age** | Teenager |
| **Lower-body type** | Dress | **Color of upper-body** | Gray |
| **Wearing hat** | No | **Color of lower-body** | White |

**Figure 1:** *Multi-view person ReID samples with associated soft biometric attributes [7]. These high-level semantic features provide viewpoint-invariant cues for robust identification.*

mid Attention (MSPA) framework [2] by implementing its core components: the Appearance Pyramid Network (APN) for visual appearance and the Attribute Pyramid Attention Network (APAN) for soft biometric features [3]. We propose to extend this fusion mechanism by explicitly accounting for the reliability of the input data. We introduce a third branch dedicated to Image Quality Assessment (IQA), based on the concept of *identifiability* [4]. The idea of this new branch is to better assess the importance of both APN and APAN features at inference time. We hypothesize that a *high-identifiability* (high-quality) image will be more reliably classified by the APN branch, and that a *low-identifiability* image will be better handled by the APAN branch. Our definition of a high-identifiability image is as follows: a query image is considered highly identifiable if its identity features remain stable under synthetic perturbations.

Specifically, we generate multiple perturbed versions of each query by simulating common surveillance degradations, including lighting shifts (HSV histogram shifts), Gaussian blur, rectangular occlusions, background color injections, and various noise types (Gaussian, salt-and-pepper). We compute an identifiability score $Q(I)$ as the average pairwise distance between the feature vectors of these perturbed images. This score serves as a dynamic weighting factor: when a query is of high quality and highly identifiable, the fusion favors the visual APN branch; conversely, for low-quality or ambiguous images, the system increases the relative influence of the more robust attribute-based APAN branch. This proposition requires a trained and reliable feature extractor model, such as the famous ResNet-50 model [5, 6].

The contributions of this work are threefold. First, we provide a complete implementation of the MSPA-based visual (APN) and attribute (APAN) branches to establish a robust baseline for the ReID task [2, 6]. Second, we propose the use of an IQA branch [4] that quantifies feature consistency under realistic synthetic perturbations, removing the need for manual quality annotations. Finally, we develop a dynamic, quality-driven weighted fusion scheme that adaptively modulates the relative importance of visual and semantic cues based on the query's reliability.

# II Related Work

## II.1 Methodological Approaches

**Multi-Task Learning and Semantic Correlations** Modern architectures often adopt a multi-task learning framework to extract identity and soft biometrics simultaneously [1, 3]. For instance, the Attribute-Person Recognition (APR) network learns identity and attributes in parallel [11]. A significant advantage of this approach is that the attribute-related loss acts as a regularizer, forcing the network to converge toward a more robust representation of identity, which improves performance even during testing phases where attribute labels may be absent [11]. However, a limitation of basic multi-task models is their potential to overlook the complex relationships between different traits.

To address this, semantic correlation modeling has been introduced. Modules such as the Attribute Reweighting Module (ARM) and the use of Convolutional Long Short-Term Memory (ConvLSTM) cells allow the system to learn statistical dependencies, such as the high correlation between "long hair" and "female" [11, 12]. This sequential learning helps filter semantic noise and updates the system's belief in its predictions. Furthermore, graph-based approaches represent attributes as nodes in a graph to discover "unseen" relationships, enhancing the model's ability to generalize to missing data or previously unobserved attribute combinations [13].

**Challenges.** Despite their effectiveness, these approaches face several practical issues: (i) multi-task setups often require dense and reliable attribute annotations, which are expensive to obtain and can be noisy or subjective; (ii) attribute labels are frequently imbalanced (rare attributes), leading to biased training and unstable gradients; (iii) learned correlations may be dataset-specific and can amplify spurious dependencies (e.g., hairstyle–gender) that do not transfer well across domains; and (iv) correlation modules (ConvLSTM/graph reasoning) increase model complexity and may be prone to overfitting when training data is limited.

**Robustness to Environmental and Appearance Variations** Environmental challenges such as low resolution and occlusion require specialized visual strategies [1]. One prominent approach is the integration of Super-Resolution (SR) modules, such as the SR-MAR model, which enhances image quality prior to

attribute extraction to maintain accuracy at long distances [14]. While effective, such methods increase the computational overhead of the system. In scenarios involving occlusion, models like Seg-DGDNet utilize semantic segmentation maps to ignore masked regions—such as surgical masks or environmental obstacles—and focus exclusively on visible biometric features [15].

Appearance variations, particularly clothing changes, represent a significant hurdle for long-term Re-ID [1]. The Masked Attribute Description Embedding (MADE) approach addresses this by using a Vision Transformer that ignores pixels associated with clothing and aligns its attention with immutable biometric traits like facial shape and body morphology [10]. While effective for cloth-changing scenarios, these multimodal approaches often require precise textual descriptions, which may not always be available in automated pipelines. In the context of this project, we explicitly focus on "non-clothes changing" tasks (Short-term ReID), where clothing remains a consistent and highly discriminative soft biometric feature [3]. Consequently, the challenges specific to long-term appearance changes fall outside the current scope of our implementation, allowing for a more specialized optimization of visual and semantic attribute fusion.

**Challenges.** Robustness-oriented modules also have clear drawbacks: (i) super-resolution pipelines can be computationally heavy and may introduce hallucinated details/artifacts that hurt ReID embeddings; (ii) segmentation-based occlusion handling depends on accurate masks, which can fail under severe blur or low resolution and propagate errors downstream; and (iii) clothes-changing or multimodal methods (e.g., requiring text) are harder to deploy at scale and may break when “immutable” cues (face/body shape) are not visible due to pose, camera angle, or privacy-preserving blur.

**Transfer Learning and Unsupervised Adaptation** The scarcity of annotated data has led to the development of semi-supervised and unsupervised strategies [1]. Techniques such as SSDAL utilize transfer learning to migrate knowledge from large, annotated datasets like PETA to unannotated target domains using triplet loss [1]. Other methods, such as Maximum Mean Discrepancy (MMD), align features statistically across domains [1]. Critical analysis suggests that while these methods significantly reduce the manual annotation burden, they currently lack the precision required for complex ReID tasks when used in isolation [1]. They are best utilized as a pre-training or refinement step within larger, more robust architectures [1].

**Challenges.** Domain adaptation remains difficult because: (i) the domain gap is often multi-factor (camera style, resolution, pose, background), so aligning global feature distributions (e.g., with MMD) may not guarantee class-conditional alignment; (ii) pseudo-labeling and self-training are sensitive to early mistakes, which can create confirmation bias and drift; (iii) transferring attribute predictors is especially fragile when attribute definitions differ across datasets (label shift); and (iv) many methods require careful hyperparameter tuning and are less stable across unseen target domains.

**Synthesis of Methodologies** The following table summarizes the primary methodologies discussed, categorized by their supervision type, key mechanisms, and inherent limitations [1].

| Model | Sup. | Mechanism & Limitation |
|---|---|---|
| APR [11] | Super. | **Joint Learning**: Efficient but requires dense annotations [11]. |
| MSPA [2] | Super. | **ConvLSTM**: Models correlations; high complexity [2, 12]. |
| SRMAR [14] | Super. | **Super-Res**: Good for distance; sensitive to artifacts [14]. |
| MADE [10] | Multi. | **ViT Masking**: Robust to clothes; needs text metadata [10]. |
| SSDAL | Semi. | **Triplet Loss**: Low annotation cost; less precise. |

**Table 1:** *Comparative synthesis of Re-ID methodologies.*

## Proposed Technical Framework and Core Architectures

Following the initial exploration of the state of the art [1], this section details the three pivotal works that constitute the technical foundation of our project. These references were selected to build a robust ReID framework capable of handling visual uncertainty through soft biometric fusion and quality-aware feature weighting.

**Core Architecture: Multi-Scale Pyramid Attention (MSPA)** The primary backbone of our implementation is the Multi-Scale Pyramid Attention (MSPA) network proposed by Khan et al. [2]. This architecture was chosen for its advanced ability to jointly manipulate the complementarity between soft bio-

metrics and visual appearance.

The MSPA model proposes a dual-branch approach: the Appearance Pyramid Network (APN) and the Attribute Pyramid Attention Network (APAN) [2]. While the APN extracts global and local identity features using mechanism attention through the Local attention Network (LTN) and multiscale information extracted to learn correlations through MFPM [9], the APAN leverages a Convolutional Long Short-Term Memory (ConvLSTM) structure [12] to maintain the semantic context among soft biometrics [3]. This specific mechanism allows the model to focus on suppressed image areas that contain crucial biometric data, making it significantly more resilient than standard ResNet-50 baselines [5]. Our work builds directly upon this fusion strategy to improve identification accuracy in non-cooperative environments.

**Feature Weighting via Image Quality Assessment (IQA)** To enhance the MSPA framework, we integrated the concept of "Identifiability" as a task-specific Image Quality Assessment (IQA), based on the research by Chen et al. [4]. In real-world surveillance, query images often suffer from varying degrees of degradation (blur, low resolution, or noise) [1].

Our methodology employs an identifiability score to dynamically weight the contributions of the APAN and APN branches. When a query image is identified as being of low quality or highly perturbed, the system adaptively re-weights the feature fusion. This prevents poor-quality visual data from degrading the final feature vector, instead prioritizing more stable soft biometric attributes when the holistic visual signal is unreliable [3]. This integration transforms a static fusion model into an adaptive, quality-aware system.

**Data Foundation: Attribute Annotation for Market-1501** Finally, the successful training of a multi-task network requires precisely labeled data for both identity and attributes. We rely on the work of Lin et al. [11], who provided the manual attribute annotations for the Market-1501 dataset [7].

This contribution is vital as it provides 27 distinct attribute labels (e.g., gender, clothing length, luggage) for the 1,501 identities in the dataset. By utilizing these high-granularity annotations, we are able to supervise the APAN network effectively, ensuring that the learned latent space reflects real-world physical characteristics. The complementarity of these ID and attribute labels is what allows our model to achieve a more discriminative and identifiable feature representation.

# III Datasets and Preprocessing

**Market-1501.** We conduct experiments on Market-1501, a large-scale person re-identification benchmark composed of cropped pedestrian images captured by multiple cameras [1, 7]. Following the standard dataset structure, we use the `bounding_box_train` split for training, `query` for query images, and `bounding_box_test` as the gallery.

**Human-annotated attributes.** We use the official Market-1501 attribute annotations as lightweight supervision for soft biometrics [3, 7, 11]: *gender; hair length; sleeve length; lower-body clothing length; lower-body clothing type; accessories (hat, backpack, bag, handbag); age group (4 classes); and upper-/lower-body clothing colors (8 and 9 binary fields, respectively, with exactly one color marked as present per identity).* You can find an exhaustive list of these soft biometrics in the table 2.

**Table 2:** *Description of attributes and representations.*

| Attribute | File Repr. | Label Values |
|---|---|---|
| Gender | gender | male (1), female (2) |
| Hair length | hair | short (1), long (2) |
| Sleeve length | up | long (1), short (2) |
| Lower-body len. | down | long (1), short (2) |
| Lower-body type | clothes | dress (1), pants (2) |
| Wearing hat | hat | no (1), yes (2) |
| Backpack | backpack | no (1), yes (2) |
| Bag | bag | no (1), yes (2) |
| Handbag | handbag | no (1), yes (2) |
| Age | age | young(1), teen(2), adult(3), old(4) |
| Upper color (8) | upblack, ... | no (1), yes (2) |
| Lower color (9) | downblack, ... | no (1), yes (2) |

**Preprocessing and data augmentation.** Across all experiments, images are resized to $384 \times 128$ and normalized using ImageNet statistics. During training, we apply random horizontal flipping ($p = 0.5$), 10-pixel padding followed by random cropping, color jittering (brightness/contrast/saturation), and random erasing ($p = 0.5$) to simulate partial occlusions. At test time (query/gallery), no augmentation is applied: images are only resized to $384 \times 128$ and normalized.

**Image transformations for IQA scoring.** Following the IQA protocol [4], we compute an identifiability score by probing the stability of a query image under

synthetic perturbations that reflect typical surveillance variations. We generate five families of perturbations: (i) lighting changes via a histogram shift on the V channel in HSV, (ii) blur via Gaussian filtering with $\sigma \in [1, 1.85]$, (iii) occlusion by overlaying a black rectangle (four sizes on each of the top, bottom, left, and right sides, i.e., 16 occluded variants), (iv) background changes by applying an oval mask to replace the image edges with a uniform color (two mask sizes and four colors: red, green, blue, yellow), and (v) noise (Gaussian, salt-and-pepper, and speckle). All perturbed images are preprocessed identically, encoded with a pretrained ResNet-50 feature extractor [5], and the identifiability score is computed as the mean pairwise Euclidean distance across the resulting feature vectors.

# IV Methodology

In this section, we present our Quality-Aware Person Re-Identification framework. We first formalize the problem and briefly review the reproduced MSPA architecture which serves as our backbone [2]. We then detail our primary contribution: the Image Quality Assessment (IQA) branch [4] and the adaptive fusion mechanism. Finally, we describe the two-stage training strategy adopted to optimize the complete network.

## IV.1 Problem Formulation

The objective of person re-identification (ReID) is to learn a mapping function $f$ that transforms a raw image $I$ into a numerical descriptor $v$. We define our training set $S$ as a collection of $N$ samples, where each sample $i$ is represented by the triplet $(I_i, y_i, a_i)$. Here, $y_i$ denotes the identity label belonging to a set of $P$ unique individuals, and $a_i$ encodes $K$ binary attributes, such that $a_i = [x_1, x_2, \ldots, x_K]$. The goal is to ensure that for any two images $I_i$ and $I_j$, the distance $d(v_i, v_j)$ is small if they share the same identity ($y_i = y_j$) and large otherwise.

To capture both the physical appearance and semantic details of a person, we utilize the Multi-Scale Pyramid Attention (MSPA) architecture [2]. This backbone decomposes the mapping function into two specialized branches. The first branch, the Appearance Pyramid Network (APN), extracts a visual feature vector $v_a$ by processing the image through multiple spatial scales. This allows the model to remain robust to changes in the person's pose or distance from the camera. Mathematically, we view this as $v_a = f_{app}(I)$, where the function $f_{app}$ focuses on global silhouettes and color distributions.

The second branch, the Attribute Pyramid Attention Network (APAN), generates a semantic feature vector $v_s = f_{attr}(I, a)$. Unlike the visual branch, this stream uses a Pyramid Attention Feature Module (PAFM) to isolate regions containing the $K$ soft-biometric attributes [3]. A key component of this branch is a Convolutional Long Short-Term Memory unit (ConvLSTM) [12], which models the relationships between attributes. For instance, it learns to refine $v_s$ by recognizing that certain traits, such as clothing types and accessories, often appear in correlation [13].

In the baseline MSPA model [2], these two components are concatenated to form a single identity descriptor $V = [v_a, v_s]$. However, in practical surveillance scenarios, the identifiability of an image depends on more than just environmental factors like blur or poor lighting. It also depends on the inherent distinctiveness of the person's appearance. For example, a person seen from the back without any accessories is visually ambiguous and harder for the visual branch to distinguish within a crowded gallery. We propose that an effective feature representation should dynamically adjust the weight assigned to $v_a$ and $v_s$ based on the identifiability of the input [4]. By quantifying how reliable a query is-considering both image quality and the presence of discriminative cues-our architecture can rely more heavily on stable soft biometrics when the global visual signal is either degraded by noise or too generic to be certain.

## IV.2 Network Architecture

The proposed framework is structured as a three-stream architecture. This design integrates the Appearance Pyramid Network (APN), the Attribute Pyramid Attention Network (APAN), and the proposed Image Quality Assessment (IQA) branch [4] into a unified pipeline. While the APN and APAN are specialized in extracting discriminative visual and semantic features, respectively, the IQA branch operates in parallel to quantify the reliability of the input. This tripartite configuration enables a dynamic fusion mechanism where the identifiability score adaptively modulates the relative influence of visual cues versus soft-biometric attributes. By converging these three specialized streams, the model produces a final identity descriptor that is optimized for both the environmental capture conditions and the inherent distinctiveness of the person's appearance. The Figure 2 is representing our architecture proposition.

**Appearance Pyramid Network (APN)** The primary objective of the Appearance Pyramid Network (APN) is to extract a robust visual representation that re-

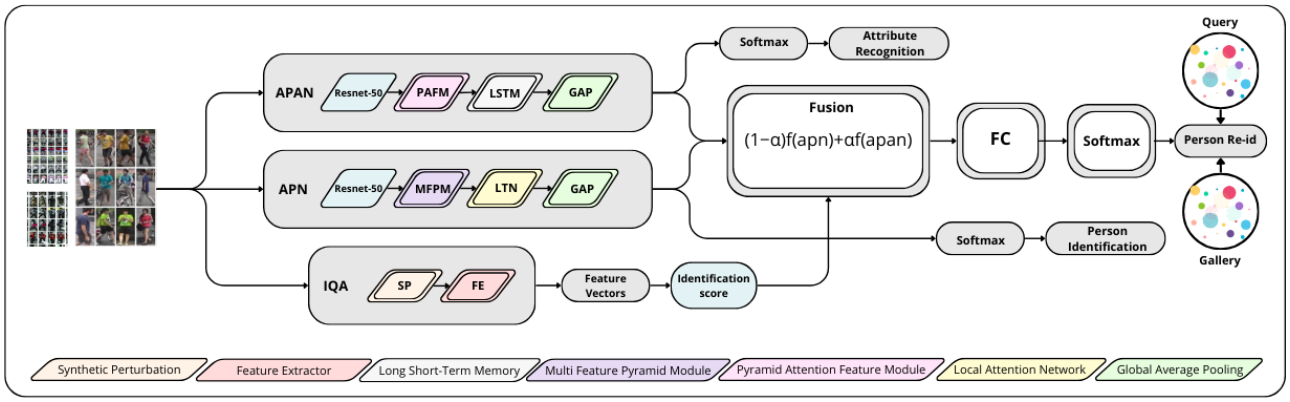


**Figure 2:** *The global architecture of our implementation of the MSPA paper [2], completed with a new IQA branch [4].*

mains invariant to scale variations and viewpoint changes. In real-world surveillance, the distance between the subject and the camera often fluctuates, making it difficult for standard convolutional layers to capture consistent patterns. The APN addresses this by encoding multi-level contextual information, ensuring that the resulting feature representation captures both fine-grained textures and the overall silhouette of the pedestrian [9].

At the core of this branch is the Multi-Feature Pyramid Module (MFPM) [2], which acts as a specialized processor for the feature maps generated by the ResNet-50 encoder [5]. The MFPM consists of four parallel convolutional paths: one traditional $1 \times 1$ convolution and three dilated convolutions with dilation rates of 2, 3, and 4. By using varying dilation rates, the network can expand its receptive field to capture wider spatial dependencies without increasing the computational overhead or losing resolution. To complement these local features, a global context path utilizes Global Average Pooling (GAP) and bilinear interpolation to capture the spatial interdependencies of the entire image.

Inside the APN, these diverse feature maps are integrated through a hierarchical decoder. This decoder performs channel-wise summation and applies a series of $1 \times 1$ convolutions to aggregate the multiscale information into a unified structure. Bilinear interpolation is used during this stage to upsample features and maintain spatial alignment across the pyramid. The final output of the branch is the visual feature vector $v_a$. This vector serves as a dense numerical descriptor of the pedestrian's global appearance, providing the primary identity signal that will later be weighted by the IQA branch.

**Soft-Biometric Attribute Branch (APAN)** The Attribute Pyramid Attention Network (APAN) [2] is designed to extract high-level semantic cues, known as soft biometrics [3], which provide a viewpoint-invariant complement to raw visual features. While the APN focuses on global appearance, the APAN is specialized in identifying localized traits—such as gender, clothing type, or the presence of a backpack—that remain consistent even when environmental conditions degrade the image quality.

The architectural flow begins with the Pyramid Attention Feature Module (PAFM) [2]. To account for the varying sizes and aspect ratios of human body parts, the PAFM processes features from the ResNet-50 encoder [5] through three parallel convolutional scales ($3 \times 3$, $5 \times 5$, and $7 \times 7$). These multiscale features are integrated with global context information via a pixel-wise attention mechanism. By multiplying global feature descriptors with the pyramid-derived attention maps, the module accurately isolates significant biometric regions while suppressing irrelevant background noise.

A distinguishing feature of this branch is the integration of a Convolutional Long Short-Term Memory (ConvLSTM) unit [12]. Unlike standard CNNs that treat attributes as independent labels, the ConvLSTM processes semantic features sequentially to model their natural correlations (e.g., the statistical relationship between hair length and gender) [11–13]. The unit uses a memory cell and a series of gates—input, forget, and output—to maintain dependencies across the set of features. This sequential learning ensures that the network's final belief about one attribute is informed by the presence of others. The output of this branch is the semantic feature vector $v_s$, which represents a structured, context-aware summary of the person's biometric profile.

## Proposition: the Quality-Aware Adaptive Fusion

Traditional architectures, such as the baseline MSPA framework [2], employ a static fusion of the Appearance Pyramid Network (APN) and the Attribute Pyramid Attention Network (APAN) at inference. While MSPA uses a weighting parameter $\lambda$ to balance loss

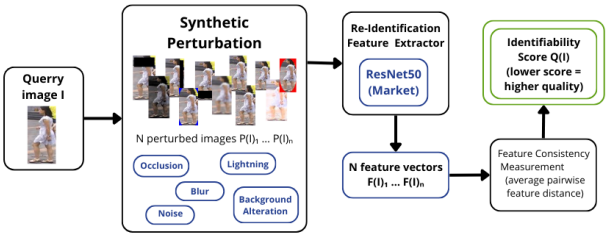


**Figure 3:** *Architecture of the IQA branch illustrating the consistency-based scoring mechanism [4].*


**Figure 4:** *A few examples showing the best and the worst images for the IQA scoring branch [4].*

functions during training, this value is unused during inference. Such static weighting fails to account for the varying reliability of feature extractors under diverse environmental conditions. We address this by introducing an Image Quality Assessment (IQA) branch [4] that computes an identifiability score at inference time, allowing the model to modulate feature importance based on the objective stability of the input.

**Hypotheses and Theoretical Framework** Our approach is grounded in two primary hypotheses regarding feature stability. First, the Reliability Shift posits that as image quality degrades, global visual signals in the APN become increasingly stochastic. Conversely, the APAN architecture—utilizing 12 prediction heads to supervise 27 attributes—is hypothesized to maintain higher semantic stability. We contend that the APAN will outperform the APN specifically in contexts involving low-identifiability images. While the APN relies on a holistic feature representation that may falter when confronted with non-distinct silhouettes or generic clothing colors, the APAN's multi-head configuration increases the probability of isolating subtle, localized cues. Consequently, even under significant degradation, the APAN is more likely to capture these discrete semantic traits than the APN is to produce a discriminative global representation. Second, the Loss-Ratio Correspondence suggests that the optimal weighting factor $\alpha$ should reflect the empirical reliability of each branch observed during training. Specifically, branches yielding lower relative loss for a specific image profile should be assigned proportionally higher influence during the fusion process.

The validations or invalidations of our hypothesis are measured by two criteria: first, achieving superior mAP and Rank-1 performance over the MSPA baseline; second, validating that low-identifiability images (quantified by high identifiability score) correlate with an increased relative reliability of the APAN branch over the APN.

**Model-Centric Identifiability** We define "identifiability" as a model-centric metric rather than a subjective human judgment. Human perception and deep neural networks respond differently to image degradation; for instance, a vision model may extract stable feature vectors from motion-blurred inputs by identifying specific chromatic distributions or anatomical ratios that are non-intuitive to human observers. Conversely, high-resolution images with high class ambiguity (e.g., a person seen from behind with a special posture) represent low-discriminative sets of features despite their clarity, as shown in the Figure 4. To quantify this [4], we utilize a ResNet-50 backbone pretrained on Market-1501 [7] as a feature extractor [5, 6]. For each query image $I$, we generate 34 modified versions $\{I'_1, \ldots, I'_{34}\}$ via isolated, non-cumulative synthetic perturbations, namely blur, lightning shifts, occlusions, background changes and additon of noise. The identifiability score $Q(I)$ is defined as the mean pairwise Euclidean distance between the 35 resulting feature vectors, representing the original input plus the 34 perturbations:

$$Q(I) = \frac{2}{N(N-1)} \sum_{i<j} d(F(I)_i, F(I)_j) \qquad (1)$$

where $N = 35$. A low $Q(I)$ indicates high representation stability, while a high score signals an unstable feature representation.

**Dynamic Weighting and MLP Optimization** To map the raw score $Q(I)$ to a weighting factor $\alpha \in [0, 1]$, we employ a lightweight Multi-Layer Perceptron (MLP) followed by a sigmoid activation function. The MLP is just a single perceptron, but this block of the network could well be a little bit larger in order to grasp the complexity of the relation between the identifiability score and the APN and APAN losses. The MLP is trained via a supervised protocol based on the cross-branch performance ratio. Using ground-truth identities available during training, we compute the relative reliability of each branch for every sample. Let $L_{apn}$ and $L_{apan}$ represent the identity losses for the respective branches. The target weight $\hat{\alpha}$ is defined as:

$$\hat{\alpha} = \frac{L_{apan}}{L_{apn} + L_{apan}} \tag{2}$$

By optimizing the MLP using cross-entropy against this ratio, the network learns to predict optimal fusion weights based solely on the input's identifiability score. The mapping is optimized such that a low $Q(I)$ yields a high $\alpha$, favoring the APN, while a high $Q(I)$ produces a lower $\alpha$, prioritizing the APAN. The final identity descriptor $V_{final}$ is constructed via weighted concatenation:

$$V_{final} = [\alpha \cdot v_a, (1-\alpha) \cdot v_s] \tag{3}$$

In a learned feature space, the network produces feature vectors that represent identity-related information. Scalar multiplication by $\alpha$ attenuates the relative influence of a branch while keeping the same feature representation structure. This mechanism ensures the final fully connected layer prioritizes the most stable feature source for each specific query, effectively transferring the model's self-assessed reliability into a robust inference-time weighting scheme.

## Final Layer for Final Prediction

Once the MLP layer has converted the identifiability score from a scalar to a weight, both the APN and APAN branches return their respective feature maps. In both cases, these feature maps are $1024 \times 1$ feature vectors, which are re-weighted according to Equation 3. To merge the outputs from both branches, we employ a final fully-connected (FC) layer. This layer identifies and compresses the concatenated $2048 \times 1$ feature vector into a single $1000 \times 1$ feature vector. Unlike the initial concatenation, this process effectively merges the features from both branches. This feature vector serves as the final descriptor for the individual in the query image. While this representation is sufficient for inference, a Softmax layer is appended at the end of the network for supervised training; this stage is bypassed during the inference phase.

## Training Strategy

The optimization of the proposed architecture is conducted via a 2-phase training strategy [6]. This approach ensures that the APN and APAN branches develop specialized feature representations before being integrated into a unified descriptor via the fusion MLP and the final fully-connected layer.

**Stage 1: Branch-Specific Optimization** In the initial phase, the APAN branch is trained independently to stabilize it's respective latent space. This branch focuses exclusively on the extraction of soft biometrics [3]. To manage the 27 detectable attributes, the branch utilizes 12 specialized prediction heads by grouping some semantics related attributes in multivalued branches. This multi-head configuration is designed to account for spatial and semantic variations; for instance, distinct heads are allocated to the upper and lower body to capture independent color distributions and textures. The global loss for the APAN branch, $L_{apan}$, is defined as the summation of the individual losses across all 12 heads.

$$L(x_i) = -\frac{1}{S}\sum_{i=1}^{S} \log \frac{e^{W_{m_{b^c}}(a_i^q + b)}}{\sum_{j=1}^{K} e^{W_{m_j}(a_i^q + b)}} \tag{4}$$

$$L_{apan} = \sum_{i=1}^{L} L\left(a_i^q\right) \tag{5}$$

where $b_i$ denotes the person identity (ID) of the $i$-th pedestrian image, and $W_m$ and $b$ are the corresponding weight matrix and bias term. The input images are processed in mini-batches of size $S$.

By pretrainning the branch APAN during the first stage, APAN learns to localize and identify fine-grained attribute-level semantics before global task so when it will be integrated in the global ReiD models, it will not slow down and mislead the trainning process of the global model by injecting widespread attribute features.

**Stage 2: Integrated Fusion and Joint Fine-Tuning** Once the branch APAN exhibits convergence and reliability, the network enters the second stage of training. This phase introduces the MLP weighting block and the final fully-connected (FC) fusion layer into the optimization loop. The MLP is trained to predict the optimal weight $\alpha$ by minimizing the cross-entropy against the target reliability ratio $\hat{\alpha}$:

$$L_{\text{MLP}} = \frac{1}{N}\sum_{i=1}^{N} \left[(1-\alpha)L_{\text{apn}} + \alpha L_{\text{apan}}\right] \tag{6}$$

$$L_{apn} = \sum_{i=1}^{H} L\left(d_i^l\right) + \sum_{i=1}^{V} L\left(a_i^q\right) \tag{7}$$

(H) and (V) indicate the horizontal and vertical numbers of stripes, respectively.

Globally,in the second training phase, the model is optimized end-to-end by jointly learning (i) the two feature extraction branches (APN and APAN) and (ii) the fusion components driven by image quality score.

Given an input batch, we perform two forward passes in parallel: the APN branch produces identity logits $\text{logits}_{\text{apn}}$ and a visual representation $v$, while

the APAN branch outputs attribute logits $\text{logits}_{\text{apan}}$ and an attribute representation $a$. The fusion block then maps the scalar identifiability score to the mixing coefficient $\alpha \in [0,1]$ , projects both feature vectors to a common embedding space, and constructs the fused descriptor via weighted concatenation. The resulting feature vector is processed by the subsequent fully-connected layers (part-level FC, feature-vector head, and global classifier) to produce global identity logits $\text{logits}_{\text{global}}$.

Training relies on multiple complementary objectives: (1) an APN identity loss (cross-entropy on $\text{logits}_{\text{apn}}$), (2) an APAN attribute loss (a combination of binary cross-entropy and multi-class cross-entropy across the attribute heads), (3) a global identity loss (cross-entropy on $\text{logits}_{\text{global}}$), and (4) a fusion-specific loss computed from the fused descriptor. For the latter, the classifier weights are detached to ensure that gradients update only the fusion parameters (including the $\alpha$-prediction module), thereby isolating the learning signal to the quality-aware mixing mechanism.

In practice, we apply four independent backpropagation steps per mini-batch, each with its own optimizer: APN parameters are updated using the APN identity loss, APAN parameters using the attribute loss, the global head using the global identity loss, and finally the fusion/IQA parameters using the fusion-specific loss.The aim is to give to each parts of the global network the possibility to learn from each their proper errors in trainning.

# V Experimental Results

This section presents experimental results for our Quality-Aware Fusion Model (QAFM), assessing both overall ReID accuracy and whether the identifiability-driven IQA branch enables query-adaptive fusion between the visual APN and the soft-attribute APAN representations. Experiments are conducted on the Market-1501 benchmark [7] using its standard train/test split and evaluation protocol. We report conventional ReID performance measures (mAP, Rank-1, Rank-5 and Rank-10) and analyze how the learned fusion weight $\alpha$ adapts to query degradation. We compare our approach against representative baselines and discuss the results in terms of both accuracy and robustness to degradation.

## Experimental Setup

This subsection describes the hardware environment, implementation choices, and the training/evaluation protocol used to validate the proposed quality-aware fusion strategy.

**Implementation Details** The full architecture is implemented in *PyTorch*. Training is performed on a single NVIDIA RTX 4500 Ada Generation GPU with 24 GB of VRAM. While our method follows the overall MSPA design [2], we re-implemented the complete pipeline in order to integrate the identifiability-driven IQA module natively and to ensure consistent data handling across the APN, APAN, and IQA branches.

**Image Transformation Pipelines** We employ three distinct image processing streams, depending on the stage and the branch:

1. **Global model training (data augmentation).** To improve robustness to common surveillance variations, all input images are resized to $384 \times 128$ and augmented using: We apply random horizontal flipping ($p = 0.5$), 10-pixel padding followed by random cropping, color jittering (brightness/contrast/saturation), and random erasing ($p = 0.5$) to simulate partial occlusions.
2. **Global model inference.** During evaluation (both query and gallery), no augmentation is applied. Images are only resized to $384 \times 128$ and normalized.
3. **IQA branch perturbations.** Following the identifiability protocol [4], we generate 34 synthetically perturbed variants for each original image. These perturbations are applied directly to the PIL image before any tensor-level preprocessing: We consider five perturbation families: lighting shifts of the HSV value channel with offsets $\{-90, -45, 45, 90\}$, Gaussian blur with $\sigma \in \{1, 1.5, 1.85\}$, border occlusions via black rectangles (from 5% to 20% of the image size), background changes using an oval mask that replaces borders with a uniform color (two mask sizes; colors red, green, blue, and yellow), and noise (Gaussian, speckle, and salt-and-pepper). Each perturbed image is then resized to $384 \times 128$ and normalized before being processed by the IQA backbone.

Across all pipelines, images are normalized using ImageNet statistics: $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ [5].

**Optimization Strategy** We use the Adam optimizer with a base learning rate of $3.5 \times 10^{-4}$. The learning schedule consists of a linear warmup over the first 10 epochs followed by cosine annealing up to 120 total epochs. To stabilize training, the APN and APAN branches are frozen during the initial 10 epochs; after unfreezing, their learning rates are reduced by a factor of 0.1.

The training objective combines multiple complementary losses: label-smoothing cross-entropy for

identity classification (with $\epsilon = 0.1$), BCEWithLogitsLoss for attribute prediction, and a triplet loss with margin 0.3 to structure the feature space.

**IQA Module Protocol** The identifiability score is computed as the mean Euclidean distance between the feature vector of the original image and the feature vectors of its perturbed variants, extracted by a ResNet-50 feature encoder pretrained on Market-1501. The resulting score is normalized (empirically, $\mu = 11.109$ and $\sigma = 0.658$) and then fed to a lightweight MLP ($1 \to 32 \to 1$) with a Sigmoid activation to produce the fusion weight $\alpha \in [0, 1]$.

## Datasets and Evaluation Protocol

**Market-1501** We conduct experiments on Market-1501 [7], a standard benchmark for person re-identification captured by multiple non-overlapping cameras in a realistic surveillance setting. We use the official training/testing split and follow the standard query–gallery evaluation protocol defined by the dataset.

**Evaluation Metrics** Performance is reported using the conventional ReID metrics: mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) accuracies at Rank-1, Rank-5, and Rank-10. mAP summarizes retrieval quality across the full ranked list, while CMC measures the probability that at least one correct match appears within the top-$k$ retrieved gallery images.

**Comparaison with State-of-the-Art (SOTA)** To contextualize the effectiveness of our quality-aware fusion strategy, we compare QAFM against representative state-of-the-art person ReID methods on Market-1501 using the standard CMC accuracies (Rank-1/5/10) and mAP. Table 3 reports the published results for prior approaches alongside our implementation.

# VI Analysis and Discussion

Overall, QAFM achieves competitive performance, with Rank-1 accuracy of 93.05% and mAP of 83.34%, confirming that quality-aware fusion can provide strong retrieval accuracy on Market-1501 [7]. Compared to earlier attribute-aware or attention-based baselines, QAFM yields a clear improvement in both Rank-1 and mAP, indicating that incorporating an explicit reliability signal helps stabilize matching under challenging conditions. However, QAFM remains below the best-reported numbers in Table 3, suggesting that our current implementation prioritizes robustness and identifiability over maximizing benchmark performance [2].

**Table 3:** *Re-identification performance comparison with state-of-the-art methods on the Market-1501 dataset.*

| Model | R-1 | R-5 | R-10 | mAP |
|---|---|---|---|---|
| PCSL [16] | 87.0 | 94.8 | 96.6 | 69.4 |
| SSPR [17] | 75.2 | – | – | 52.6 |
| EEA [18] | 87.9 | – | – | 71.0 |
| PL-Net [19] | 88.2 | – | – | 69.3 |
| MGTS [20] | 84.8 | – | – | 67.0 |
| Deep-ReID [21] | 86.2 | 92.4 | 96.1 | 65.4 |
| CDP [22] | 74.5 | 83.8 | – | 51.4 |
| MCFL [23] | 84.1 | – | – | 64.9 |
| UECIR [24] | 98.0 | 98.6 | 99.5 | 98.3 |
| LR-Net [25] | 91.4 | – | – | 79.6 |
| HCM [26] | 91.8 | 96.7 | 97.7 | 79.0 |
| UPR [27] | 93.0 | – | – | 82.4 |
| MSPA [2] | 97.7 | 98.9 | 99.7 | 98.0 |
| QAFM | 93.05 | 97.62 | 98.60 | 83.34 |

**Table 4:** *Experimental results (ablation and fusion) on Market-1501.*

| Experiment | Rank-1 | Rank-5 | mAP |
|---|---|---|---|
| QAFM | 93.05 | 97.62 | 83.34 |
| apn_only_alpha_0 | 86.43 | 94.51 | 71.24 |
| apan_only_alpha_1 | 87.00 | 94.83 | 68.10 |
| fixed_alpha_0.1 | 88.06 | 95.37 | 74.28 |
| fixed_alpha_0.2 | 90.77 | 96.82 | 78.74 |
| fixed_alpha_0.3 | 92.19 | 97.45 | 81.67 |
| fixed_alpha_0.4 | 92.70 | 97.45 | 81.91 |
| fixed_alpha_0.5 | 91.63 | 97.33 | 80.03 |
| fixed_alpha_0.6 | 90.62 | 96.85 | 77.04 |
| fixed_alpha_0.7 | 89.37 | 96.20 | 73.98 |
| fixed_alpha_0.8 | 88.51 | 95.58 | 71.40 |
| fixed_alpha_0.9 | 87.47 | 95.13 | 69.44 |

**Distribution of identifiability score (IQA score).** Fig. 5 shows that the identifiability-based identifiability score (IQA score) [4] are concentrated around $\approx 11.1$ (mean 11.13, median 11.14), with relatively limited spread (roughly from $\approx 9.1$ to $\approx 13.1$). This indicates that, on Market-1501, most queries have comparable "feature stability" under our synthetic perturbations, and only a small fraction exhibit extremely low or extremely high identifiability. In practice, such a narrow range reduces the dynamic leverage that the IQA module can exert on the fusion weight.

**Learned fusion behavior ($\alpha$).** As shown in Fig. 6, the learned weights collapse to a very narrow interval around $\alpha \approx 0.375$ (mean 0.375), far from the MSPA baseline fixed $\alpha = 0.5$ [2]. Since higher $\alpha$ corresponds to a stronger reliance on the visual APN branch,

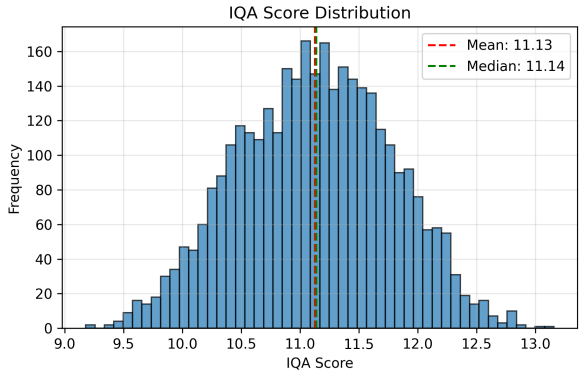


**Figure 5:** *Distribution of the identifiability-based identifiability score computed on Market-1501 queries.*

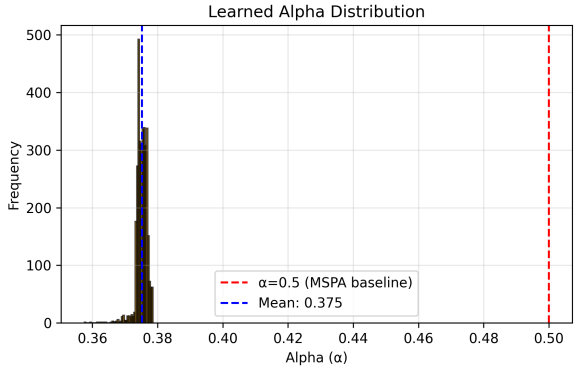


**Figure 6:** *Distribution of the learned fusion weight $\alpha$ on Market-1501 queries.*

this shift suggests that, for our training setup, the model systematically favors the soft-biometric APAN branch. Interestingly, this is consistent with the fixed-$\alpha$ ablation in Table 4, where the best fixed weights lie in the $\alpha \in [0.3, 0.4]$ range; the learned solution effectively recovers this operating point.

**Does IQA drive $\alpha$?** Fig. 7 reports a weak (slightly negative) correlation between IQA and $\alpha$ ($r = -0.090$) and an almost flat linear fit. This shows that, although the fusion module is conditioned on the identifiability score, the resulting mapping is close to constant over the observed IQA range. A likely explanation is that the IQA distribution is too concentrated (Fig. 5), making it difficult for the MLP to learn a strongly discriminative, sample-specific weighting. Consequently, the current gain of the IQA branch appears to come more from learning a robust global mixing ratio than from truly adapting to each query.

**Qualitative evidence and failure modes.** The activation visualization in Fig. 8 suggests that both branches attend primarily to identity-relevant regions on the body (upper torso, carried accessories, and lower-body cues), while suppressing most of the

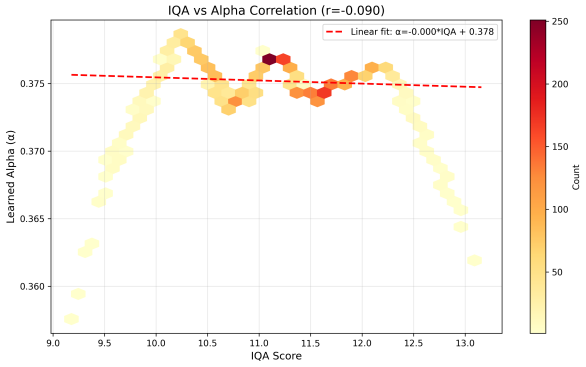


**Figure 7:** *Relationship between the identifiability score and the learned fusion weight $\alpha$.*


**Figure 8:** *Qualitative visualization (heatmap) highlighting regions that contribute most to the matching decision.*

background. This supports the intuition that the two branches provide complementary evidence: APN emphasizes appearance patterns, whereas APAN tends to focus on more semantic, localized cues [2, 3].

The retrieval examples further illustrate both the strengths and the limitations of the approach. In Fig. 9 (query ID 533, $\alpha \approx 0.374$), the correct identity dominates the top ranks, indicating that the fused descriptor is stable even under moderate blur/low resolution; the single mismatch near Rank-10 highlights the residual ambiguity caused by similar clothing colors (green upper garment) and pose variations. In Fig. 10 (query ID 352, $\alpha \approx 0.375$), the top-1 match is correct, but several distractors appear in the top-10, reflecting the challenging combination of severe motion blur and appearance similarity in dark clothing. Notably, both examples yield very similar $\alpha$ values, again consistent with the near-constant fusion behavior observed in Fig. 6–7.

Overall, these results indicate that QAFM learns an effective global fusion point (close to the best fixed-$\alpha$), but the current IQA-driven conditioning does not yet translate into strong query-adaptive mixing. Improving the sensitivity of the IQA signal (e.g., by widening the perturbation set, refining the score normalization, or learning IQA features jointly with the fusion) is a promising direction to obtain more dynamic, per-query weights.

# VII Conclusion

This work investigated quality-aware fusion of visual appearance and soft-biometric attributes for per-


**Figure 9:** *Successful example where the quality-aware fusion improves retrieval under moderate degradation.*


**Figure 10:** *Difficult example illustrating remaining failure cases under severe ambiguity/degradation.*

son re-identification in realistic surveillance conditions. Building on the MSPA framework [2], we implemented the APN (visual) and APAN (attribute) branches and introduced an additional identifiability-driven IQA module [4] to modulate the fusion weight at inference time. Experiments on Market-1501 [7] show that our approach (QAFM) achieves strong performance (Rank-1 93.05%, mAP 83.34%) and that combining visual and semantic cues is consistently beneficial compared to using either branch alone.

Our analysis further reveals that, under the current protocol, the learned fusion behaves mostly as a stable global mixture (with $\alpha \approx 0.375$) rather than a strongly query-adaptive mechanism, due to the limited variability of identifiability scores on this benchmark. Despite this, the qualitative results suggest that the fused model attends to identity-relevant regions and remains robust under moderate degradations, while severe blur and high appearance similarity remain challenging.

Future work will focus on strengthening the adaptivity of the quality signal and its coupling with fusion. Promising directions include enriching the perturbation set and calibration of the identifiability score, learning the IQA mapping jointly with the ReID backbone, and extending evaluation to more diverse datasets and cross-domain settings where image quality varies more substantially.

# References

[1] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022. doi: 10.1109/TPAMI.2021.3054775.

[2] Samee Ullah Khan, Noman Khan, Tanveer Hussain, Khan Muhammad, Mohammad Hijji, Javier Del Ser, and Sung Wook Baik. Visual appearance and soft biometrics fusion for person re-identification using deep learning. *IEEE Journal of Selected Topics in Signal Processing*, 17(3): 575–586, 2023. doi: 10.1109/JSTSP.2023.3260627.

[3] Shan Lin and Chang-Tsun Li. Person re-identification with soft biometrics through deep learning. In *Deep Biometrics*, pages 21–36. Springer International Publishing, Cham, 2020. ISBN 978-3-030-32583-1. doi: 10.1007/978-3-030-32583-1_2. URL `https://doi.org/10.1007/978-3-030-32583-1_2`.

[4] Haoyu Chen, Edward J. Delp, and Amy R. Reibman. Estimating image quality for person re-identification. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2021. doi: 10.1109/MMSP53017.2021.9733688.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[7] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[8] Manuel J. Marín-Jiménez, Shiqi Yu, Yasushi Makihara, Vishal M. Patel, Maneet Singh, and Maria de Marsico. Editorial introduction to the special issue on biometrics at a distance in the deep learning era. *IEEE Journal of Selected Topics in Signal Processing*, 17(3):539–544, 2023. doi: 11.1109/JSTSP.2023.3269211.

[9] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[10] Chunlei Peng, Boyu Wang, Decheng Liu, Nannan Wang, Ruimin Hu, and Xinbo Gao. Masked attribute description embedding for cloth-changing person re-identification. *IEEE Transactions on Multimedia*, 27:1475–1485, 2025. doi: 10.1109/TMM.2024.3521730.

[11] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2019.06.006. URL `https://www.sciencedirect.com/science/article/pii/S0031320319302377`.

[12] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf`.

[13] Xi Yang, Xiaoqi Wang, Nannan Wang, and Xinbo Gao. Address the unseen relationships: Attribute correlations in text attribute person search. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16916–16926, 2024. doi: 10.1109/TNNLS.2023.3300582.

[14] Ramin Abbaszadi and Nazli Ikizler-Cinbis. Merging super resolution and attribute learning for low-resolution person attribute recognition. *IEEE Access*, 10:30436–30444, 2022. doi: 10.1109/ACCESS.2022.3159102.

[15] Muskan Dosi, Chiranjeev Chiranjeev, Shivang Agarwal, Jyoti Chaudhary, Sunny Manchanda, Kavita Balutia, Kaushik Bhagwatkar, Mayank Vatsa, and Richa Singh. Seg-DGDNet: Segmentation based disguise guided dropout network for low resolution face recognition. *IEEE Journal of Selected Topics in Signal Processing*, 17(6):1260–1272, 2023. doi: 10.1109/JSTSP.2023.3288398.

[16] L. Qi, L. Wang, J. Huo, Y. Shi, and Y. Gao. Progressive cross-camera soft label learning for semi-supervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2815–2829, Sep 2020.

[17] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang, and N. Zheng. Semi-supervised person re-identification using multi-view clustering. *Pattern Recognition*, 88:285–297, 2019.

[18] J. Wu et al. An end-to-end exemplar association for unsupervised person re-identification. *Neural Networks*, 129:43–54, 2020.

[19] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing*, 28(6):2860–2871, Jun 2019.

[20] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai. Person search by separated modeling and a mask-guided two-stream cnn model. *IEEE Transactions on Image Processing*, 29:4669–4682, 2020.

[21] S. U. Khan, T. Hussain, A. Ullah, and S. W. Baik. Deep-reid: Deep features and autoencoder assisted image patching strategy for person re-identification in smart cities surveillance. *Multimedia Tools and Applications*, pages 1–22, 2021. doi: 10.1007/s11042-020-10145-8.

[22] Y. Ge, L. Liu, and H. Zhang. A three-stage learning approach to cross-domain person re-identification. *Applied Soft Computing*, 112: 107793, 2021.

[23] S. Zhou, J. Wang, J. Shu, D. Meng, L. Wang, and N. Zheng. Multinetwork collaborative feature learning for semisupervised person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4826–4839, Sep 2022.

[24] M. Wieczorek, B. Rychalska, and J. Dabrowski. On the unreasonable effectiveness of centroids in image retrieval. In *Proceedings of the 28th International Conference on Neural Information Processing*, pages 212–223, 2021.

[25] S. U. Khan, I. U. Haq, N. Khan, K. Muhammad, M. Hijji, and S. W. Baik. Learning to rank: An intelligent system for person re-identification. *International Journal of Intelligent Systems*, 37:5924–5948, 2022.

[26] T. Si, F. He, Z. Zhang, and Y. Duan. Hybrid contrastive learning for unsupervised person re-identification. *IEEE Transactions on Multimedia*,

2022. doi: 10.1109/TMM.2022.3174414. Early access, May 11, 2022.

[27] T. Liu, Y. Lin, and B. Du. Unsupervised person re-identification with stochastic training strategy. *IEEE Transactions on Image Processing*, 31:4240–4250, 2022.