biomimetics

MDPI

*Article*

# Bio-Inspired Enhanced Quantization-Aware Training-Based Knowledge Distillation for Person Re-Identification

**Ho Won Lee, Sung Jae Lee, Su Jin Im, Hyun Woo Song, In Ho Jeung and Kang Ryoung Park ***

Division of Electronics and Electrical Engineering, Dongguk University, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Republic of Korea; ih1541@dongguk.edu (H.W.L.); sungje0319@dgu.ac.kr (S.J.L.); sujin125917@dgu.ac.kr (S.J.I.); nasa1832@dongguk.edu (H.W.S.); inhoj@dgu.ac.kr (I.H.J.)
* Correspondence: parkgr@dongguk.edu

**Abstract**

Infrared (IR) person re-identification (Re-ID) supports surveillance in nighttime and low-light conditions, but low resolution, noise, and limited texture information hinder recognition. We propose a bio-inspired lightweight framework based on enhanced quantization-aware training-based knowledge distillation (EQT-KD). Selective feature suppression and adaptive weighting of head, upper-body, and lower-body regions provide a functional, design-level analogue of selective visual attention. Enhanced LightMBN serves as the teacher, while Extreme LightMBN is the compact student. EQT-KD adjusts branch-wise distillation strength according to the difference between student outputs before and after fake quantization, supporting stable knowledge transfer under quantization. Experiments on SYSU-MM01, DBPerson-Recog-DB1, and the near-infrared (NIR) subset of RGBNT201 yielded mean average precision values of 0.7646, 0.9883, and 0.7473 and Rank-1 accuracies of 0.8037, 0.9903, and 0.7547, respectively, under the IR-only protocol used in this study, in which each unseen test identity has one query and two gallery images. The student uses 2.35 million parameters and 0.66 GFLOPs. Inference-time measurements on desktop, NVIDIA Jetson Orin Nano, and mobile platforms indicate the potential for real-time processing. These results support selective feature processing and quantization-aware knowledge transfer as a practical approach to balancing recognition performance and computational efficiency in resource-constrained IR surveillance.

**Keywords:** Bio-inspired learning; low-resolution infrared imaging; person re-identification; quantization-aware training; knowledge distillation; edge surveillance

Check for updates

Academic Editors: Yongquan Zhou, Heming Jia and Qing Gao

Received: 23 September 2026
Revised: 3 November 2026
Accepted: 16 November 2026
Published: 18 November 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Person re-identification (Re-ID) is a task that determines whether a person captured by surveillance cameras is the same person by comparing the person across images from different cameras, and it plays a key role in various application fields, such as intelligent surveillance and unmanned security systems [1,2]. In this study, Re-ID is addressed as IR-to-IR retrieval of person images, in which the identities used for testing are not seen during training.

In particular, Re-ID based on visible light cameras shows excellent performance because it can utilize high-resolution images, but it has a limitation in that stable image acquisition is difficult in low-light environments, such as cloudy weather or nighttime [3]. To alleviate this problem, Visible–Infrared (VI) multimodal Re-ID studies that jointly use visible light (VL) and infrared light (IR) have recently been actively conducted [4]. VI-based approaches

have greatly improved performance through cross-modality alignment, adversarial feature mapping, and other methods. However, because two different types of camera sensors must be operated simultaneously, system cost increases, and the network structure becomes more complex, leading to increased computational complexity and memory requirements. As a result, their application is limited in low-cost and low-power edge-based surveillance environments [5].

Meanwhile, IR Re-ID studies that use only IR cameras, without an additional visible-light camera, have the advantage of enabling stable imaging even in nighttime and low-light environments. Here, IR includes near-infrared (NIR) images captured with active NIR illumination and thermal images that record emitted long-wave infrared radiation; both types are used in this study (Section 4.1). However, in real surveillance environments, subjects are often captured from far away from the camera, and situations frequently occur in which only IR images containing low resolution and noise can be used. To solve this problem, some studies have proposed a pipeline that performs Re-ID after super-resolution (SR), but this approach has limitations in that the training process is complex and the computational cost is high [6]. Furthermore, existing IR-based Re-ID studies have mainly been designed with a focus on performance improvement, and studies on lightweight models or quantization-based approaches remain very limited [7]. In particular, studies that compensate for performance degradation caused by quantization in low-resolution IR environments through distillation have not been sufficiently addressed. Consequently, a lightweight Re-ID model that can operate efficiently while using only low-resolution IR images as input remains an important unresolved research problem. Smart surveillance and edge computing require lightweight Re-ID models that balance recognition performance with limited computational resources and power budgets [8]. In robotic visual recognition, pruning and knowledge distillation have reduced model size and computational cost for edge deployment [9]. Such efficiency-aware design is also essential for memory-constrained platforms, including NVIDIA Jetson and mobile devices.

Selective visual attention provides a biological perspective on emphasizing relevant image information while limiting the influence of competing visual inputs [10]. From this bio-inspired perspective, the proposed framework combines selective feature suppression in the global branch with adaptive weighting of anatomically defined body regions in the part branch. Its contribution to biomimetics is an engineering realization of selective feature processing for resource-constrained infrared recognition, with knowledge distillation transferring the resulting representation to a compact quantized model. This connection is a functional design analogy; the experiments evaluate recognition and computational efficiency rather than biological fidelity. Specifically, each module implements one computational principle of selective attention. The part weight gate (PWG) implements competitive allocation of a limited processing weight among body regions, because the softmax normalization forces the head, upper-body, and lower-body weights to compete, analogous to the competition among stimuli for limited processing resources [10]. Top DropBlock (Top-DB) and the unimportant feature remover (UFR) suppress the most and least salient spatial rows, respectively, so that recognition cannot rely on a single dominant region or on uninformative background regions, analogous to the suppression of distractors. The width-split channel (WSC) structure processes the left and right halves of the person independently, analogous to the separate processing of the two visual hemifields. These correspondences are design-level analogies; the modules are not intended as models of the underlying neural mechanisms.

Against this background, this study proposes a lightweight IR Re-ID framework based on enhanced quantization-aware training-based knowledge distillation (EQT-KD) for IR

surveillance environments containing low resolution and noise. The contributions of this paper are as follows.

- In the global branch of the teacher model, we propose a URT (UFR + RFR + Top-DB) module by adding the unimportant feature remover (UFR) and random feature remover (RFR) to Top DropBlock (Top-DB). Top-DB suppresses overfitting to specific patterns by removing rows with high feature values. The UFR module removes rows with low feature values, such as floor regions or noise, in low-resolution infrared images, thereby preventing those regions from being used. Through this process, the model is guided to use high feature values, which is intended to make it more robust to background changes and noise in IR images. In addition, during the training process, the RFR module randomly removes feature values, which encourages the model to learn representations that do not depend on individual feature elements and are therefore better suited to noisy low-resolution infrared images. To compensate for the lack of detailed information caused by low resolution, each remaining feature is further strengthened to have a more robust representation capability. Even if some feature map elements are randomly removed from blurred boundaries caused by low resolution, the model is trained to compensate for the boundaries by combining the remaining neighboring elements.
- In the part branch of the teacher model, we propose a part weight gate (PWG) module. The PWG module calculates importance by using a shared 1 × 1 Conv block for each of the three parts, head, upper body, and lower body. The channel reduction by the shared convolution is intended to make the importance estimate less sensitive to noise in low-resolution IR images. Then, by summarizing the entire region of each part into a single scalar, overfitting to a specific channel or detailed pattern is prevented. The calculated importance is normalized by softmax and multiplied by each part feature. Because these weights are computed from the input, the network can adjust the contribution of each body part according to the image rather than treating all parts uniformly.
- In the channel branch of the teacher model and the student model, we propose a width-split channel (WSC) structure. The WSC structure separates a feature map into left and right sides and trains each side with an independent 1 × 1 Conv. In low-resolution IR images, noise that occurs only on one side, such as occlusion or thermal noise, can appear. By learning the left and right sides as separate representations, each supervised by its own classifier, the discriminative cues of each side are learned independently and are both retained in the final descriptor.
- We propose a loss weighting (KLW) module based on Kullback–Leibler (KL) divergence. The KLW module takes the difference between the classifier logits of the student before and after applying fake quantization (FQ), namely the quantization error ($q_e$), for each classifier, inputs it into a single linear layer to calculate weights, and applies them to the classifier-wise KL loss of the student model, thereby adjusting the strength of distillation for each classifier according to the $q_e$. The KLW module is excluded from the target of quantization-aware training (QAT) and is trained in a full-precision state. Through this, additional quantization error inside the KLW module can be prevented, and the strength of distillation for each classifier can be adjusted more stably. In the inference stage, the KLW module and the teacher model are not used, and the model parameters of the student model are quantized, resulting in low memory usage and reduced computational cost.

The rest of this paper is organized as follows. Section 2 introduces the related works relevant to this study. Section 3 describes the details of the proposed model. Section 4

presents the experimental results and their analysis, Section 5 presents the discussions, and Section 6 presents the conclusion of this paper.

## 2. Related Works

This section reviews existing Re-ID studies on VL, VI, and IR images. For each modality, methods requiring high computation cost achieve high accuracy with large backbones and complex modules, but their computational and memory requirements limit their use on edge devices and in real-time surveillance; therefore, lightweight methods are reviewed separately.

### *2.1. VL Image-Based Methods*

#### 2.1.1. Requiring High Computation Cost

Recently, in VL-based Re-ID, studies using large-scale backbones and complex network structures have been actively conducted to solve various real-world problems, such as complex environmental changes, occlusion, and clothing changes. Shape-aligned supervision (SEAS) [11] introduced shape-aligned supervision using human body shape information and improved Re-ID performance through feature learning aligned with the human body structure. It uses implicit 3D body-shape representations as pixel-level supervision during training rather than as an additional input, so the shape information adds cost mainly to training and does not increase the inference cost; its computational cost at inference is therefore that of its appearance-based Re-ID network. The multi-feature fusion [12] method strengthened the discriminative capability of pedestrian representations through a multi-branch structure that simultaneously utilizes global and local features. By combining features at different levels, it aimed to utilize information useful for pedestrian identification more richly. Training generalized person re-identification on a mixture of data (Remix) [13] proposed a generalized person re-identification training strategy that mixes different data distributions, aiming to improve generalization performance against various environmental changes. Through mixed-data training, it aimed to learn representations robust to domain changes. Synergistic interaction framework (SIF) [14] proposed a perception-aware optimization and interaction-based framework considering multiple occlusion situations. By adjusting feature interactions according to the degree of occlusion and scenario changes, it aimed to improve recognition performance in complex occlusion environments. Feature alignment with adversarial learning for clothing-changing person re-identification (FA-ReID) [15] introduced an adversarial learning-based feature alignment strategy to learn stable person representations under clothing-changing situations. It focused on learning clothing-invariant features by using human parsing information and an attention mechanism.

#### 2.1.2. Requiring Low Computation Cost

Recently, in the field of VL image-based Re-ID, studies on lightweight models for reducing computational complexity and the number of parameters have been actively conducted with the goal of application to edge devices and real-time surveillance environments. To maintain identification performance while securing efficiency compared with existing heavyweight models, various approaches to lightweight backbone design and efficient feature structures have been proposed. Multi-scale semantic and detail extraction network (MSDENet) [16] proposed a structure that efficiently combines semantic and detail information at multiple scales and learns discriminative feature representations even in a lightweight network. By separating and integrating semantic information and detailed information within a shallow network structure, it improved the representation capability of the lightweight model. The structural redundancy reduction-based method [17] proposed

a lightweight Re-ID backbone that efficiently utilizes multi-scale information through a combine and expand training strategy and structural re-parameterization. While maintaining rich representations in the training stage, it secured computational efficiency by converting the structure into a simple branch structure in the inference stage.

Attention knowledge-aided distillation lightweight network (ADLN) [18] introduced an attention-based self-distillation strategy and proposed a lightweight structure that transfers information from deep layers to shallow layers within a single network. Through this, it reduced the number of model parameters and inference delay while maintaining competitive performance. Multi-scale feature enhancement network (MSFE-Net) [19] combined multi-scale feature enhancement and an attention mechanism to perform feature learning robust to complex backgrounds and various visual changes. While maintaining a lightweight structure, it improved representation capability through multi-granularity feature interaction. The random color dropout-based lightweight Re-ID method [20] utilized a lightweight backbone such as Omni-scale network (OSNet) [21] and introduced a data augmentation technique that alleviates dependence on color information. By complementing spatial and channel information through a self-attention module, it achieved stable performance even in a lightweight network.

Overall, existing VL image-based lightweight model-based Re-ID studies have attempted to achieve a balance between computational efficiency and performance through lightweight backbones, efficient multi-scale designs, attention, and augmentation techniques. However, these methods are mainly optimized for VL image information, and dependence on complex visual cues still exists. This supports the need for lightweight Re-ID studies in a single-modal IR image environment robust to illumination changes.

### *2.2. VI Image-Based Methods*

#### 2.2.1. Requiring High Computation Cost

Recently, in the field of VI Re-ID, studies based on large-scale backbones and complex network structures have been actively conducted to alleviate the large modality gap between visible and infrared. Dual modality-aware alignment (DMA) [22] proposed an alignment framework that simultaneously considers the features of VL and IR images. By performing modality-aware alignment in a dual-stream structure, it strengthened the learning of common representations between modalities and cross-modal feature consistency. High-order hierarchical middle-feature learning network (HOH-Net) [23] proposed a hierarchical middle-feature learning structure using high-order relationships to progressively alleviate the modality gap in VI Re-ID. Through fine-to-coarse hierarchical feature interaction and attention-based alignment, it improved the discriminative capability of middle-level representations. The bidirectional modality information interaction [24] method introduced a bidirectional information exchange mechanism between visible and infrared features. Through iterative modality interaction and feature aggregation, it performed representation learning that simultaneously strengthened inter-modality and intra-modality information. CLIP-driven semantic discovery network (CSDN) [25] introduced high-level semantic information based on contrastive language-image pre-training (CLIP) into VI Re-ID and compensated for visual differences between VI modalities in the semantic space. Through semantic prompts and a feature integration strategy, it aimed to simultaneously achieve cross-modal alignment and representation generalization. Frequency domain nuances mining (FDNM) [26] addressed the VI Re-ID problem from the perspective of the frequency domain and proposed cross-modality feature learning using amplitude and phase components. Through salience-guided phase enhancement and amplitude nuances mining modules, it performed representation learning differentiated from spatial domain-centered approaches.

#### 2.2.2. Requiring Low Computation Cost

Recently, in the field of VI Re-ID, studies on lightweight models that improve computational efficiency while considering the large domain gap between modalities and enabling application to edge devices and real-time surveillance environments have gradually increased. These studies aim to minimize performance degradation by using lightweight backbones, efficient training strategies, or knowledge transfer techniques. Task-oriented pretrained lightweight neural network (TOPLight) [27] proposed a task-oriented pretraining strategy based on lightweight backbones such as MobileNetV3, ShuffleNetV2, and GhostNet. Through fake domain loss and specialized data augmentation, it strengthened modality-shared feature learning between VI modalities and alleviated the problem of limited feature capacity by using a fine-grained dependency reconstruction module. The multimodal data augmentation (MDA)-based approach [28] proposed a method for improving VI Re-ID performance using a lightweight dual-stream structure based on Resnet-18. Through data augmentation techniques such as modality masking, multimodal patch mixing, and multimodal random erasing, it strengthened complementary feature learning between VI modalities and robustness to corrupted data. Dual alignment knowledge distillation (DAKD) [29] proposed a framework that improves the VI Re-ID performance of a lightweight student model by using knowledge distillation. Through instance-level and class-level alignment, it simultaneously learned modality-invariant and modality-specific representations, and effectively alleviated the distribution gap between VI modalities through cutswap-based data augmentation.

These lightweight VI methods still assume a multimodal structure including visible information and involve structural complexity or additional training cost, which motivates a simple and efficient lightweight Re-ID model for single-modal IR images.

### *2.3. IR Image-Based Methods*

#### 2.3.1. Requiring High Computation Cost

Recently, IR image-based Re-ID studies have been conducted with the goal of stable recognition performance in nighttime and low-light environments. Due to the characteristics of IR images, in which VL information is limited, many studies have approached this problem by using large-scale backbones and complex network structures for strong representation learning. The existing study by Zhang et al. [30] first systematically addressed the Re-ID problem in nighttime environments and proposed the KnightReID dataset. This study improved recognition performance in nighttime scenarios by constructing a pipeline that combines a denoising network and an existing Re-ID model while considering the noise characteristics of IR images. Original and attention-guided densenet-based ensemble network (OADE-Net) [7] proposed an ensemble structure that combines original DenseNet and attention-guided DenseNet to compensate for insufficient feature information in IR images. By using multiple networks based on DenseNet-161 and attention modules, it strengthened the discriminative feature learning of infrared images and achieved competitive performance even in a single infrared modal environment. Weak saliency ensemble network (WSE-Net) [31] proposed an ensemble network introducing a weak saliency mechanism to alleviate the influence of background noise in infrared-based Re-ID. By combining a ConvNeXt-series backbone and an attention-guided model and using grouped convolution and feature reduction techniques, it improved recognition performance in complex infrared environments.

#### 2.3.2. Requiring Low Computation Cost

IR image-based Re-ID studies have mainly been conducted with a focus on improving recognition performance in nighttime and low-light environments. However, most existing

studies have attempted to solve the problem of insufficient information in IR images by using large-scale backbones or complex network structures, and approaches with lightweighting as the main objective are limited.

### *2.4. Quantization and Knowledge Distillation for Efficient Models*

Quantization reduces the memory footprint and computational cost of neural networks by representing weights and activations with low-bit integers. Post-training quantization converts a trained full-precision model without retraining, whereas quantization-aware training (QAT) simulates quantization during training with fake-quantization operations so that the network can adapt to the quantization error. Because low-precision models often lose accuracy, knowledge distillation (KD) from a full-precision teacher has been combined with quantization, and quantization-aware KD objectives have been proposed for convolutional networks [32,33] and Transformer encoders [34]. In person Re-ID, efficient models have mainly been obtained through lightweight architectures, such as OSNet [21] and the lightweight models reviewed in Sections 2.1.2, 2.2.2 and 2.3.2 [16–20], whereas quantized Re-ID models, particularly for low-resolution IR images, have rarely been studied.

Against this background, this study proposes EQT-KD, a lightweight Re-ID method for a single-modal IR environment. Based on a teacher model robust to low-resolution IR images, teacher–student knowledge distillation is performed. In addition, by applying EQT-KD to the student model, this study proposes a lightweight Re-ID model that simultaneously considers computational efficiency and recognition performance.

## 3. Proposed Method

### *3.1. Overall Procedure of Proposed Method*

Figure 1 shows the overall procedure of the proposed EQT-KD framework for low-resolution IR image-based person Re-ID in consumer edge surveillance environments. The proposed framework consists of two stages: training with knowledge distillation and quantized inference. First, low-resolution IR images are simultaneously fed into Enhanced LightMBN (teacher), an improved version of lightweight multi-branch network for Re-ID (LightMBN) [35], and Extreme LightMBN (student), which is a more lightweight model. Enhanced LightMBN is designed to extract robust features from low-resolution and noisy IR images, while Extreme LightMBN is designed to reduce computational complexity and memory usage for edge and mobile deployment. During the training stage, the teacher model is frozen and provides knowledge to the student model through EQT-KD. The proposed framework uses several core modules to improve robustness and efficiency. The URT block enhances robust global feature learning, the WSC structure performs left–right split-channel learning to learn the two halves of the body independently, the PWG module adaptively adjusts the importance of body parts, and the KLW module adjusts the strength of distillation according to quantization error. In particular, EQT-KD performs logit distillation based on KL divergence for each classifier and uses the output difference before and after fake quantization in the student model to estimate quantization error. Based on this quantization error, the KLW module controls the classifier-wise distillation strength, enabling the student model to follow the teacher model more effectively under the QAT environment. After training, the teacher model and the KLW module are no longer used, and the trained student model is quantized into Quantized Extreme LightMBN. In the inference stage, Quantized Extreme LightMBN performs query–gallery matching for person Re-ID with low computational complexity and low memory usage. This enables efficient deployment in consumer edge surveillance systems, such as Jetson-like edge devices and mobile embedded environments, while supporting real-time inference with reduced power consumption.

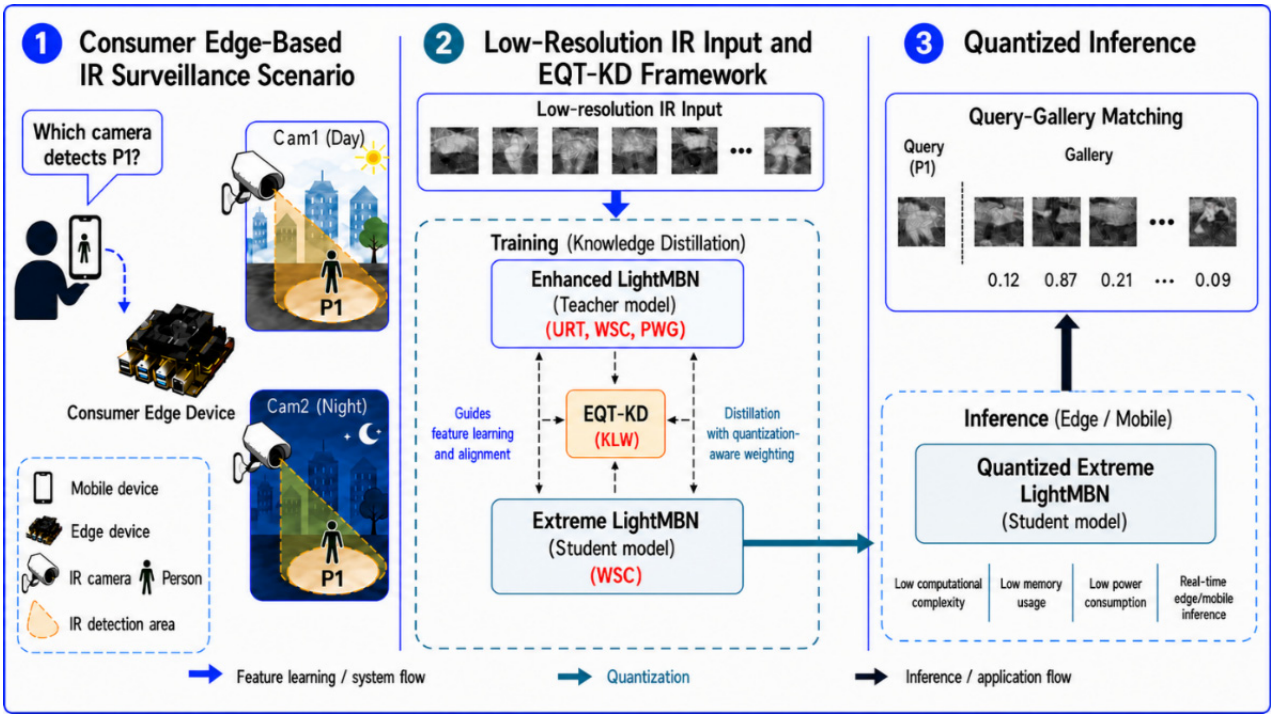



**Figure 1.** Overview of the proposed EQT-KD framework for low-resolution infrared-light person re-identification in consumer edge surveillance applications.

### *3.2. Enhanced LightMBN (Teacher Model)*

In this study, Enhanced LightMBN, which improves the existing LightMBN structure, is used as the teacher model to perform stable feature extraction in IR surveillance environments containing low resolution and noise. Figure 2 shows the overall structure of the proposed Enhanced LightMBN. As the backbone, OSNet ($\beta = 1.0$), with width multiplier β = 1.0 corresponding to the default channel setting, is used [21], and the model is composed in parallel of three branches, global, channel, and part, to learn mutually complementary features. In the global branch, to reduce dependence on background and noise in low-resolution environments, we propose the URT module, which adds UFR and RFR to the existing Top-DB. In the channel branch, the WSC structure is introduced to respond to noise concentrated on one side. In addition, PWG is applied to the part branch to adaptively weight the head, upper-body, and lower-body features.

#### 3.2.1. URT Block

The global branch learns global features based on whole-body information and aims to reduce dependence on background and noise in low-resolution IR images. In the existing LightMBN, Top-DB, which removes high feature values during the training process, was used to induce robust global feature learning. In this study, this is extended, and the URT block shown in Figure 3 is applied to the global branch. The UFR module removes rows with low feature values in the feature map. Through this, dependence on regions that do not contribute to discrimination, such as floor regions or noise, is suppressed, and the model is guided to focus on regions with high feature values, which is intended to make the model more robust to background changes and noise; this is a design hypothesis supported only by the ablation in Section 4.3.2. The RFR module randomly removes some elements of the feature map during the training process; this regularizes the model so that it does not rely on individual feature elements, which is beneficial for low-resolution IR inputs containing Gaussian noise. Even if some feature map elements are randomly removed

from blurred boundaries of low-resolution images, the network learns to compensate for boundaries by combining the remaining neighboring elements. After the URT block, pooling suitable for each output is applied, and the model is trained through an independent batch normalization neck (BNNeck) [36]. In implementation terms, the output of the global branch is first passed through an additional OSNet bottleneck block, producing a feature map $F \in R^{512\times24\times8}$. For each sample, the row activation is computed as $a_i = \max_j \|F_{:,i,j}\|_2^2$ after L2 normalization over all spatial positions. Top-DB sets to zero the round$(0.33 \times 24) = 8$ rows with the largest $a_i$, and UFR sets to zero the 8 rows with the smallest $a_i$. RFR multiplies F element-wise by a Bernoulli mask M with $P(M = 0) = 0.33$; no rescaling is applied to the remaining elements. The Top-DB feature is max-pooled, whereas the UFR, RFR, and unmasked features are average-pooled, and each pooled feature is passed to its own BNNeck and classifier. The masks are applied only during training. The phrase that the remaining features are "further strengthened" refers to the intended effect of training under these masks, in which the classifiers must rely on the unmasked elements; no explicit amplification of the remaining elements is performed. The classifiers of the UFR and RFR features provide auxiliary supervision during training only and are not included in the retrieval descriptor of the teacher.

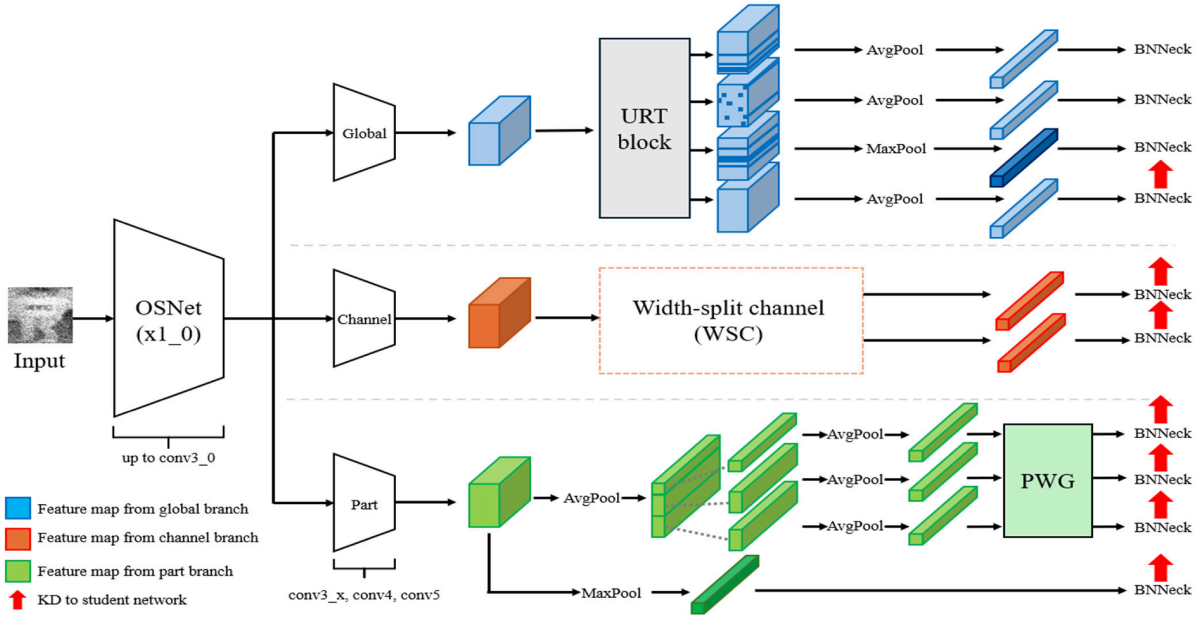



**Figure 2.** Architecture of Enhanced LightMBN.

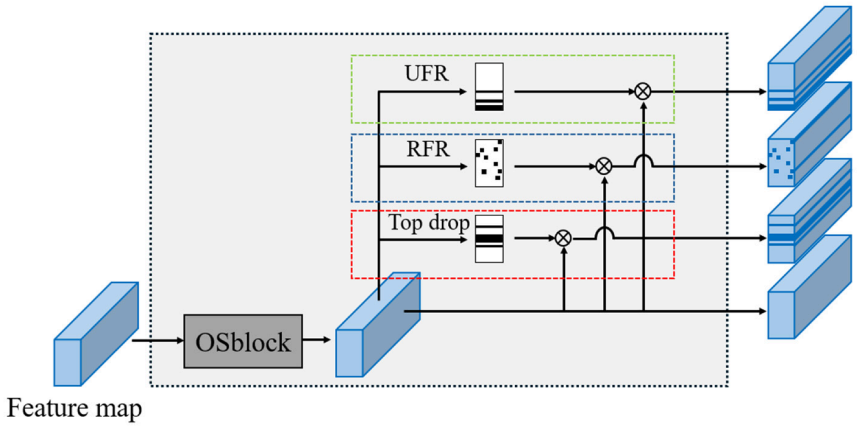



**Figure 3.** URT block.

3.2.2. WSC

The channel branch of the existing LightMBN has a limitation in that it cannot consider noise imbalance between the left and right sides because it integrates and learns all channels as one global representation. The channel branch of Enhanced LightMBN is based on the WSC structure of Figure 4, which separates the feature map into left and right sides. In low-resolution IR images, cases occur in which occlusion or noise appears only on one side of the human body. Reflecting this characteristic, the WSC structure divides the feature map into the left and right sides and trains each side through an independent 1 × 1 convolution and BNNeck. Because each side is supervised by its own classifier, the discriminative cues of the left and right halves are learned independently, and both are retained in the retrieval descriptor. Formally, the channel-branch feature map is first average-pooled to a 1 × 2 spatial map (512 × 1 × 2), and its two columns give two vectors, $c_L$, $c_R \in R^{512}$, corresponding to the left and right halves of the image width. Each vector is processed by its own 1 × 1 convolution, batch normalization, and ReLU, which are not shared between the two sides, and is then passed to its own BNNeck and classifier. WSC contains no explicit gating; both sides always contribute to the training loss and to the retrieval descriptor.

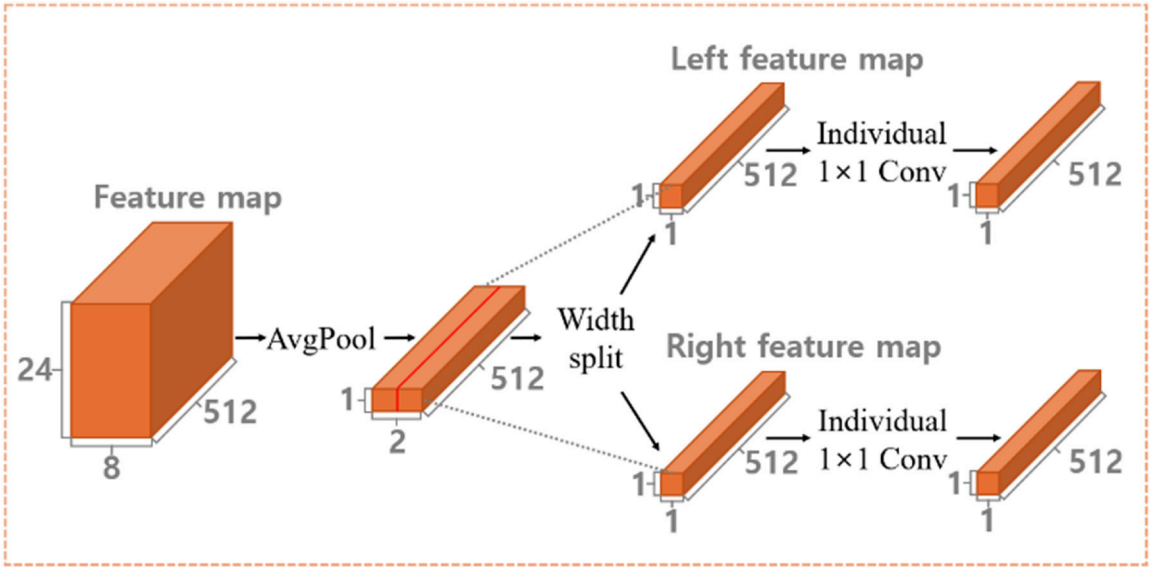



**Figure 4.** Architecture of WSC.

3.2.3. PWG

The part branch of the existing LightMBN divides the human body into two parts, upper body and lower body, and learns the features of each part with the same importance. Due to this, when a specific body part is occluded or noise occurs in a low-resolution infrared image, that part can be reflected with the same weight as other parts, which may lead to performance degradation. In contrast, in Enhanced LightMBN, to reflect the structural characteristics of the human body more finely, the feature map is divided into three parts, head, upper, and lower, and we propose the PWG module shown in Figure 5, which dynamically adjusts the importance of each part. PWG prevents overfitting by using a shared 1 × 1 convolution for each part, and the channel reduction is intended to make the importance estimate less sensitive to noise in low-resolution IR images. Then, the relative importance between parts is calculated through softmax, and this is multiplied by each part feature and reflected. In this way, the gate assigns input-dependent weights to the three parts, allowing the network to emphasize the parts that are more informative for identification. As a result, the part branch of Enhanced LightMBN can learn part-

based representations centered on structurally important body parts even in low-resolution infrared environments. Formally, the part-branch feature map is average-pooled into 12 horizontal stripes, which are grouped into the head (stripes 1–2), upper body (stripes 3–7), and lower body (stripes 8–12) and averaged into $p_{head}$, $p_{upper}$, $p_{lower} \in R^{512}$. A shared gating function $g(\cdot)$, consisting of a 1 × 1 convolution from 512 to 128 channels, ReLU, and a 1 × 1 convolution from 128 channels to one output, produces a score for each part, and the weights are obtained as $w_k = \exp(g(p_k))/\Sigma_{k'} \exp(g(p_{k'}))$. Each part feature is then rescaled as $w_k\, p_k$ before its BNNeck. Because the softmax weights sum to one, their average is 1/3; however, each BNNeck begins with a bias-free 1 × 1 convolution followed by batch normalization, so a common scale factor is removed by the normalization, and only the relative weighting between samples and parts affects the subsequent features. Because the stripe boundaries are fixed, the three regions may not correspond exactly to the head, upper body, and lower body for truncated crops or large pose changes.

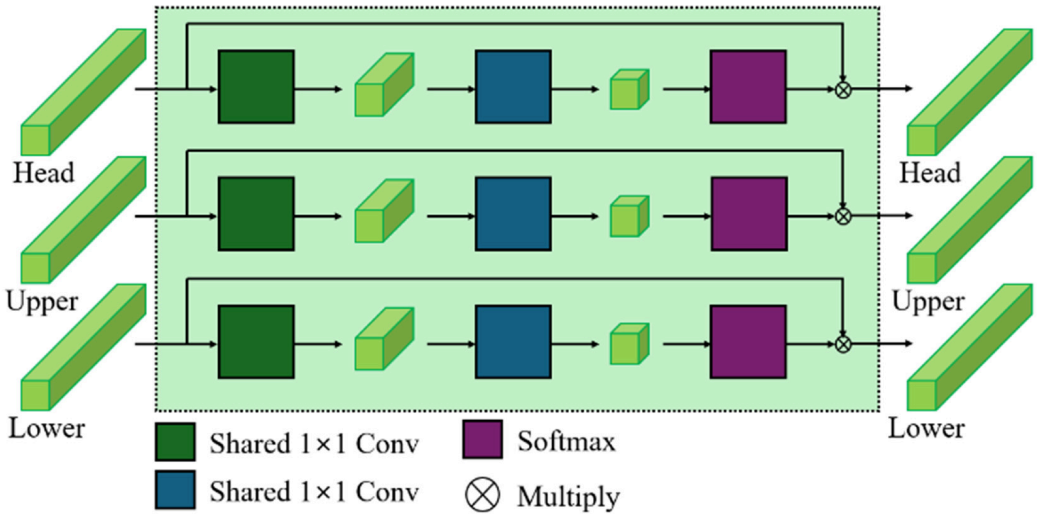



**Figure 5.** PWG.

### *3.3. Extreme LightMBN (Student Model)*

In this study, for efficient inference in low-resolution infrared surveillance environments, we designed Extreme LightMBN, as shown in Figure 6, as a lightweight student model based on the existing LightMBN. Extreme LightMBN is configured to maintain the advantages of the multi-branch structure while minimizing computational complexity and memory through structural simplification and quantization consideration. The existing LightMBN uses OSNet ($\beta = 1.0$) as the backbone and learns features at various levels through a multi-branch structure composed of global, part, and channel branches. This structure has a limitation in that it is designed on the premise of floating-point-based operations and does not sufficiently consider the efficiency of computational complexity and memory. In contrast, Extreme LightMBN greatly reduces computational complexity and memory through a lightweight backbone and an EQT-KD-based training structure, which enhances QAT with knowledge distillation.

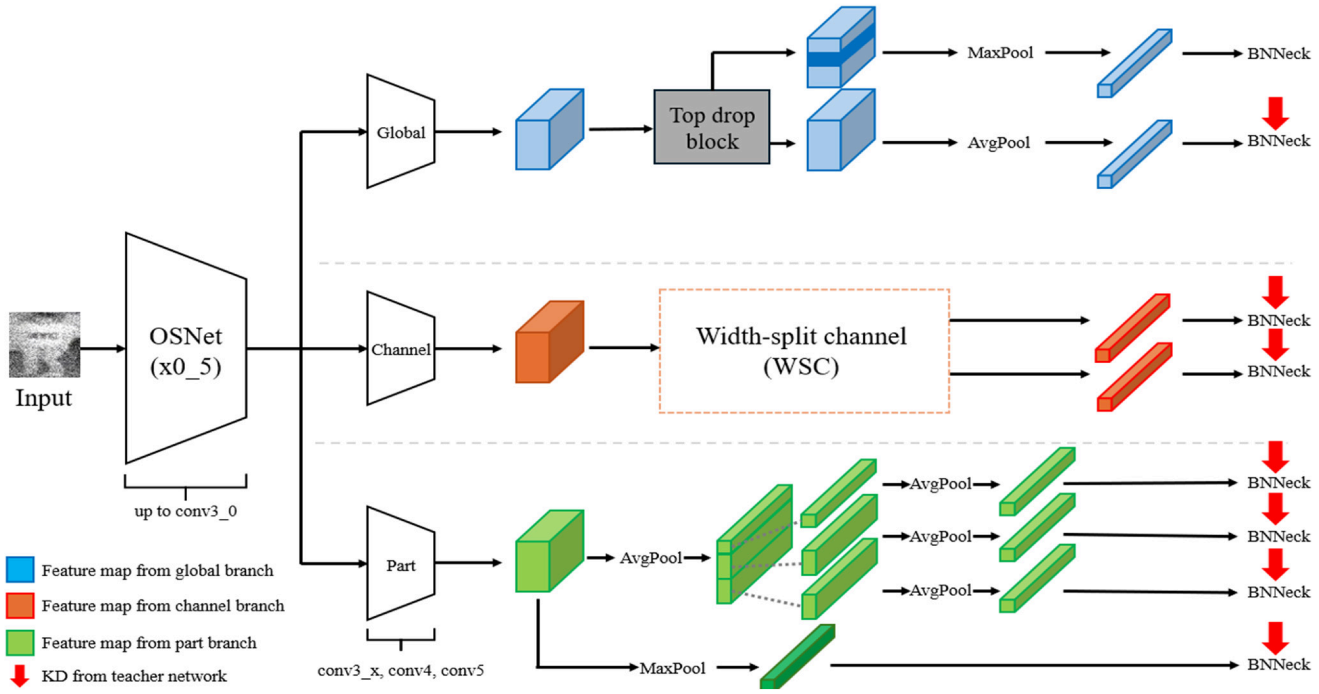



**Figure 6.** Architecture of Extreme LightMBN.

#### 3.3.1. Backbone Lightweighting

While the existing LightMBN uses OSNet ($\beta = 1.0$), Extreme LightMBN uses OSNet ($\beta = 0.5$), with width multiplier $\beta = 0.5$, which has half the number of channels of OSNet ($\beta = 1.0$), as the backbone [21], thereby reducing the number of parameters and computational complexity. The reduction in the number of channels simultaneously reduces convolutional computational complexity and the number of parameters. In addition, the feature dimension delivered to the subsequent branch structure and BNNeck module is also reduced, and thus the computational complexity and memory usage of the entire model are greatly reduced. Of the 2.35 M parameters of the student on SYSU-MM01, approximately 1.83 M belong to the embedding-extraction network (backbone, branches, and BNNecks), and the remaining parameters belong to the classifier heads, which are used only during training, and to KLW. For comparison, the teacher has 8.97 M parameters, of which 1.26 M belong to its classifier heads.

#### 3.3.2. WSC

The channel branch of Extreme LightMBN has the WSC structure. In low-resolution infrared images, cases occur in which noise and occlusion are concentrated on one side of the human body. Reflecting these asymmetric characteristics, WSC is also introduced into Extreme LightMBN to independently learn left and right features, so that the features of the two sides are learned separately and both are retained in the descriptor. In addition, Extreme LightMBN maintains the WSC structure, enabling knowledge distillation (KD) from the teacher during training.

#### 3.3.3. Part Branch

Extreme LightMBN learns partial features by dividing the feature map into three regions, head, upper, and lower, in the part branch. This is to secure relatively stable features even in low-resolution infrared environments by reflecting the structural characteristics of the human body. In the existing teacher model, the importance of each part is dynamically adjusted through PWG, but PWG is not applied to Extreme LightMBN. The reason is to

prevent an increase in the parameters of the student model and maintain the lightweight structure, and to simplify the quantized computation by removing the additional gating operations; a student with PWG was not trained, so this choice is a design simplification rather than an experimentally verified requirement. Therefore, Extreme LightMBN maintains the part division structure itself but is designed as a concise structure without additional modules. Because PWG is applied only in the teacher, the part-level logits of the teacher are computed from gate-weighted part features, whereas those of the student are computed from unweighted part features. Since distillation is applied to the classifier outputs, the student is trained to reproduce the part-level class distributions of the teacher, which already reflect the PWG weighting, rather than the gate weights themselves. Consequently, the student cannot adaptively re-weight the body parts at inference, which is a trade-off for its lower parameter count and simpler quantized computation.

#### 3.3.4. EQT-KD

Extreme LightMBN applies enhanced QAT in the training stage to use integer-based operations (INT8) in the inference stage. In general, when quantization is applied after training a model, performance degradation can occur.

To alleviate this, this study designed the model to simulate the quantization effect during the training process. The network assumes FQ, to which quantization is applied in the training stage, and through this, it adapts in advance to distribution changes that may occur in an integer operation environment. This approach helps minimize performance degradation even when actual quantization is applied during inference. In a low-resolution IR environment, the student model suffers from performance degradation due to two factors: reduced representation capability caused by lightweighting and distribution distortion caused by quantization. To solve these simultaneously, this study proposes EQT-KD, which integrates enhanced QAT and KD, as shown in Figure 7. This study differs from existing QAT in that it adjusts the strength of distillation by considering $q_e$ generated inside the student model.

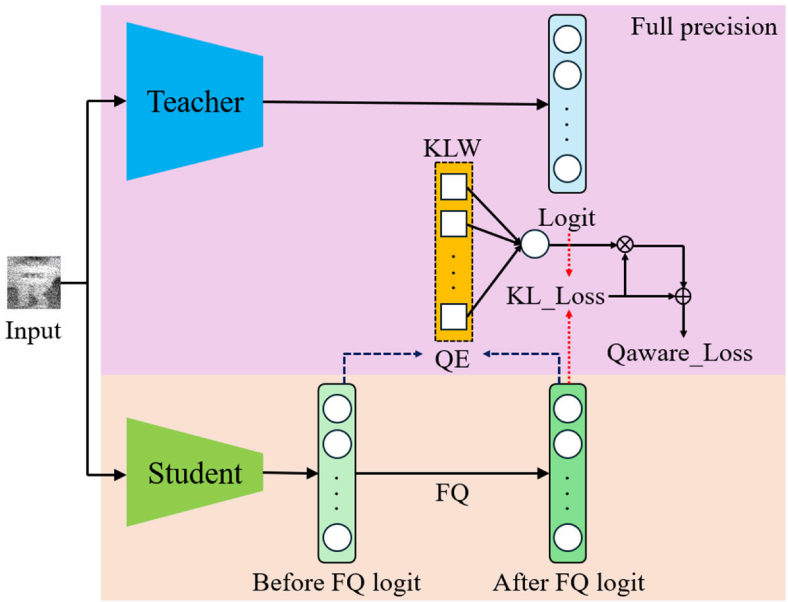



**Figure 7.** EQT-KD network.

Motivation

In the student model to which QAT is applied, two types of logits are generated from each classifier.

$$Z_s^{pre} : \text{output before fake quantization} \quad (1)$$

$$Z_s^{post} : \text{output after fake quantization} \quad (2)$$

The two outputs are generated from the same feature, but distribution distortion occurs due to integer approximation in the FQ process. As a result, the logit distribution of each branch may differ, and this distortion can lead to performance degradation. The student model of this study has a multi-branch structure, like the teacher model, and has global, part, and channel branches. Since each classifier has different structural characteristics and representation ranges, $q_e$ may also differ. Therefore, in this study, the $q_e$ generated in the logit is measured to adjust the strength of distillation.

Classifier Logit Expression

The logit output of each classifier of the student is defined as follows.

$$Z_s^{pre} \in R^{B \times H \times C} \quad (3)$$

$$Z_s^{post} \in R^{B \times H \times C} \quad (4)$$

Here, $B$ is the batch size, $H$ is the number of classifiers (7 in this study), and $C$ is the number of classes. The logit output of each classifier of the teacher model is defined as follows.

$$Z_t \in R^{B \times H \times C} \quad (5)$$

This structure enables knowledge distillation between corresponding classifiers while maintaining the multi-branch structure. Specifically, Enhanced LightMBN has ten classifiers: four in the global branch (attached to the unmasked global feature and to the features processed by Top-DB, UFR, and RFR in the URT block), four in the part branch (the max-pooled part-branch feature and the head, upper-body, and lower-body features), and two in the channel branch (the left and right WSC features). Extreme LightMBN has eight classifiers: two in the global branch (the unmasked global feature and the Top-DB feature), the same four part-branch classifiers, and the same two channel-branch classifiers. The $H = 7$ classifiers used for distillation are paired by their position in this structure: global, part-branch global, head, upper body, lower body, left channel, and right channel. The classifiers attached to the drop-based features of the global branch (Top-DB in both models, and UFR and RFR in the teacher) receive stochastically or adaptively masked inputs during training and have no one-to-one counterpart in the other model; therefore, they are trained with the identity classification loss only and are not used for distillation. For retrieval, the eight BNNeck features of the student are concatenated into a 2048-dimensional descriptor.

Definition of $q_e$

The $q_e$ of the logit for each classifier is defined as follows.

$$q_e = \left| Z_s^{post} - Z_s^{pre} \right| \quad (6)$$

Here, the L1 difference is used to measure the quantization error between $Z_s^{post}$ and $Z_s^{pre}$. This value numerically represents the degree to which the logit is affected by FQ. The pre- and post-quantization logit difference is an observable signal of how strongly FQ changes the output of each classifier and is used as the input for adapting the corresponding distillation weights. In the implementation, both logits are computed from the same QAT features and fake-quantized classifier weights, and they differ only in the fake quantization applied to

the classifier output; thus, $q_e$ reflects the effect of fake quantization on the classifier output rather than the difference between a full-precision model and the quantized model. Because this raw L1 difference depends on the scale of the logits, it is not a calibrated measure of the quantization sensitivity of each branch. Specifically, the difference in Equation (6) is taken element-wise in absolute value, so that $q_e \in R^{B\times H\times C}$ contains, for each sample b and classifier h, the C-dimensional vector $|Z_s^{post} - Z_s^{pre}|$; its sum over the class dimension corresponds to the L1 difference.

KLW

To adjust the strength of distillation based on $q_e$, this study introduces a KLW module using a lightweight linear layer. Here, KLW denotes KL-divergence loss weighting: the module does not compute a KL-divergence term itself, but predicts the weights applied to the classifier-wise KL-divergence terms. The KLW module consists of a single linear layer with a weight vector of dimension C and no bias. This layer is shared by all H classifiers and is applied to the C-dimensional $q_e$ vector of each sample b and classifier h, yielding $\omega_{\text{learned}} \in R^{B\times H}$. Thus, the class dimension is reduced by the linear layer, whereas the batch dimension is retained, so that each sample–classifier pair receives its own weight. $q_e$ is not detached from the computational graph; therefore, the gradient of the weighted distillation loss propagates both to the KLW layer and, through $q_e$, to the student logits. The KLW layer uses the default PyTorch initialization and is optimized jointly with the student by the same Adam optimizer.

$$\omega_{\text{learned}} = Linear(q_e) \tag{7}$$

A scalar weight is calculated for each classifier, and the final KL loss weight $\omega_h$ for each classifier is defined as follows.

$$\omega_h = 1 + \delta \cdot \omega_{\text{learned}} \tag{8}$$

$q_e$ can have a large variation depending on the classifier. If the weight is used as it is, the KL loss can increase greatly, making training unstable. By applying scaling with $\delta = 0.01$, the basic strength (=1) is maintained, while the correction according to $q_e$ is intended to remain at a fine-adjustment level, thereby stabilizing training. No positivity or normalization constraint is applied to the learned weights, so this scaling does not bound them mathematically; however, in the trained models, the weights remained between 0.987 and 0.996 for all classifiers, that is, KLW changed the distillation weights by at most about 1.3%. In this study, separate KLW modules are not assigned to each classifier, and one KLW module is used. That is, one linear layer receives the $q_e$ of each classifier as input and simultaneously predicts the weights for each classifier. Because the parameters of KLW are shared, the $q_e$ vectors of all classifiers are mapped to weights by the same function; each classifier is processed independently, and the $q_e$ vectors of different classifiers are not mixed. To calculate $q_e$ more precisely, QAT is not applied to KLW, and the KLW module is not executed during inference (it is disabled in the evaluation code rather than deleted), so it adds no inference computation. This module enables stable EQT-KD without increasing inference complexity. All reported results were obtained with the student model LMBN_n_osnet_x0_5_student_scaling of the public repository, which implements Equation (8); the other student variants in the repository, such as one with a sigmoid-based weight, were not used.

### *3.4. Quantized Extreme LightMBN*

After training is completed, in the inference stage, the teacher model and the KLW module are not used, and the student model is quantized. As a result, the student model is converted into INT8-based Quantized Extreme LightMBN.

The final model maintains the multi-branch structure, while convolution and linear operations are performed as INT8 integer operations. Through this, the following advantages can be obtained.

- Reduction in model memory usage.
- Inference on CPU-based edge devices through reduction in computational cost.
- Reduction in power consumption and improvement in the possibility of real-time inference.

Therefore, Quantized Extreme LightMBN operates as a lightweight and low-computational-cost model while maintaining competitive recognition performance.

### *3.5. Loss*

This study uses the following loss function to simultaneously perform identity classification, metric learning, and quantization-aware distillation between the teacher model and the student model.

$$L_{total} = \lambda_{ce} L_{ce} + \lambda_{ms} L_{ms} + \lambda_{kd} L_{kd} \tag{9}$$

In this experiment, the loss weights were set to the configuration with the highest Rank-1 accuracy in the sensitivity analysis on the test set in Section 4.3.2, namely $\lambda_{ce} = 0.5$, $\lambda_{ms} = 0.5$, and $\lambda_{kd} = 0.25$. In addition, in the multi-branch structure, the loss for each branch of the student model is calculated separately and then summed to be reflected in the final loss. Specifically, the identity classification loss is computed for every classifier of the student (eight classifiers) and summed, the MS loss is computed for the global, Top-DB, and part-branch global features and summed, and the distillation loss is computed for the seven distilled classifiers as described in Section 3.5.3.

#### 3.5.1. Identity Classification Loss

Identity classification loss is defined based on cross-entropy and uses label smoothing [37]. In this implementation, the label smoothing coefficient is set to $\epsilon = 0.1$. In addition, the target distribution is softened as follows.

$$y_c^{LS} = \begin{cases} 1 - \epsilon + \epsilon/C, & c = y \\ \frac{\epsilon}{C}, & c \neq y \end{cases} \tag{10}$$

Finally, it is defined as

$$L_{ce} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{c=1}^{C} y_{i,c}^{LS} log p_{i,c} \tag{11}$$

#### 3.5.2. Multi-Similarity Loss

To improve the separability of embedding in Re-ID, multi-similarity (MS) loss is applied. MS loss strengthens the discriminative capability of feature embedding by simultaneously considering hard positives and hard negatives based on the cosine similarity of positive-pair and negative-pair samples within a batch. In this study, the margin is set to 0.8. The detailed terms follow the definition of MS loss [38].

$$L_{ms} = \frac{1}{B} \sum_{i=1}^{B} (L_i^{pos} + L_i^{neg}) \tag{12}$$

#### 3.5.3. Quantization-Aware Distillation Loss

Distillation is performed based on KL divergence, which aligns the distribution of student logits by using the teacher logits $Z_t$ as soft targets. In this study, $Z_s^{pre}$ and $Z_s^{post}$, which are output by the student during the QAT process, have different roles: both are

used to compute $q_e$ as the input of KLW in Equation (7), whereas the KL-divergence term in Equation (13) is computed only between the teacher logits and the post-FQ student logits. $T$ is set to 1.5.

$$L_{kd} = \sum_{h=1}^{H} \omega_h KL\left( softmax\left(\frac{Z_t}{T}\right) \| \, softmax\left(\frac{Z_s^{post}}{T}\right)\right) \cdot T^2 \quad (13)$$

Here, $H$ is the number of classifiers, and $\omega_h$ is the weight considering the quantization error in the classifier, as shown in Equation (8). The KL divergence is computed in the direction $\mathrm{KL}(\mathrm{p_t} \parallel \mathrm{p_s})$, where $\mathrm{p_t}$ and $\mathrm{p_s}$ denote the temperature-softened softmax distributions of the teacher logits and the post-FQ student logits, respectively. For each sample and classifier, the KL divergence is summed over the C classes and multiplied by the corresponding weight $\omega_\mathrm{h}$, which is computed separately for each sample. The weighted values are then averaged over the B samples in the batch, summed over the H classifiers, and multiplied by $\mathrm{T}^2$.

### *3.6. Differences from Existing QAT-Based KD Methods*

Previous studies have combined QAT and KD to improve reduced-precision models [32,33], and KD objectives have also been investigated for quantized Transformer encoders [34]. EQT-KD focuses on multi-branch IR Re-ID: it measures the change in each classifier's logits before and after fake quantization and uses KLW to adjust the corresponding KL-divergence loss weights. KLW remains in full precision during training and, like the teacher, is not used during inference, so the adaptive weighting introduces no additional inference cost. The effect of combining QAT with distillation in this framework is examined in Section 4.3.2, including comparisons with uniform KL-KD and with the FP32 student trained with KD, whereas a direct comparison with the methods in [32–34] is left for future work.

## 4. Experimental Results with Analyses

### *4.1. Experimental Data and Setup*

In this study, DBPerson-Recog-DB1 [39], SYSU-MM01 [40], and RGBNT201 [41] were used to evaluate single-modality IR image-based Re-ID performance. Because these three datasets can be freely used for academic purposes according to the instructions of the dataset providers, informed consent is exempted for our research. For DBPerson-Recog-DB1, only thermal images captured in outdoor environments were used, whereas IR images captured in indoor and outdoor environments were used for SYSU-MM01. DBPerson-Recog-DB1 [39] is the dataset that is also distributed as RegDB and is referred to as Reg in our code. For SYSU-MM01, the IR images of cameras 3 and 6 were used, which covered 494 identities in our copy of the dataset; the official identity lists of the cross-modality protocol (491 identities) were not used. For RGBNT201, only the near-infrared (NIR) images of the 141 identities of the official training set (train_141) and the 30 identities of the official test set were used (171 identities); the 30 identities of the official validation set were not used. No visible-light images or multimodal combinations were included. In addition, to simulate real surveillance environments, low-resolution (LR) images with a size of 64 × 64 pixels were generated using Equation (14) [42].

$$I_{LR} = (I_{HR} * k_{blur}) \downarrow_s + n \quad (14)$$

A high-resolution (HR) image $I_{HR}$ was processed by applying the convolution operation of Gaussian blur $k_{blur}$. Then, downscaling to 64 × 64 pixels was performed. Finally, to reflect the noise environment of real situations to some extent, Gaussian noise $n$ was additionally applied to generate the LR image $I_{LR}$. When applying Gaussian blur, $\mu = 0$ and $\sigma = 3$ were set, and when applying Gaussian noise, $\mu = 0$ and $\sigma = 10$ were set. The same

degradation was applied to all images of the training, validation, query, and gallery sets, with fixed degradation parameters, and the resulting 64 × 64 images were stored offline. During training and testing, these LR images were resized to 384 × 128 pixels using bilinear interpolation, which is the input size used by all compared models except MaxViT and Swin Transformer V2 (Section 4.3.3). The LR images of DBPerson-Recog-DB1 and SYSU-MM01 were generated with the protocol of our previous study [8]: the HR images were resized to 256 × 256 pixels with Lanczos interpolation, blurred with a Gaussian filter ($\sigma = 3$), downscaled by a factor of four to 64 × 64 pixels, and corrupted with additive Gaussian noise ($\sigma = 10$ on the 8-bit intensity scale). The LR images of RGBNT201 were generated with the same degradation parameters (Gaussian blur with $\sigma = 3$, 64 × 64 pixels, and Gaussian noise with $\sigma = 10$). In all datasets, the degraded images are grayscale, and a fixed noise realization was used for each image.

In addition, as shown in Table 1, experiments were conducted using 2-fold cross validation for each dataset. Based on the number of identities, the dataset was divided in half into the train set and the test set, and approximately 20% of the images of each training identity were assigned to the validation set (two of the ten images per identity for DBPerson-Recog-DB1, the rounded-up 20% for RGBNT201, and six images per identity for SYSU-MM01, i.e., 18.8% of the training images). In addition, one query image and two gallery images were randomly selected once from each test identity, and the same fixed split was used for all models. The query and gallery images are different images, and no camera or temporal constraint was imposed on their selection; because the query images were assigned a separate camera label during data preparation or by the data loader (label 1 for the SYSU-MM01 queries, whose IR cameras are 3 and 6, and labels 0 and 1 for all RGBNT201 query and gallery images, respectively), the camera-based filtering of the evaluation code removed no gallery images, and all queries (247, 206, and 86/85 per fold) were evaluated. This label does not indicate the original camera of the query image; therefore, the evaluation does not ensure cross-camera matching, and the MD5 checks described below exclude identical files but not adjacent or highly similar frames. The 171 identities used from the NIR data of RGBNT201 were divided into folds of 85 and 86 identities. For all three datasets, the identity lists of each fold confirmed that the training and validation identities were disjoint from the test identities, and MD5 checks found no identical image file shared across the training, validation, query, and gallery sets. Our proposed method was implemented using PyTorch (version 2.5.1). Our experiment was conducted on a desktop computer equipped with an Intel ® Core™ i5-4690K @ 3.5 GHz (4 cores) CPU, 16 GB RAM, and an NVIDIA GeForce RTX 2070 graphics processing unit (GPU) with 8 GB VRAM [43].

This evaluation protocol differs from the standard SYSU-MM01 protocol [40], which evaluates visible–infrared cross-modality retrieval in the all-search and indoor-search modes. In this study, only IR images were used for both the query and gallery sets, all images were degraded to low resolution, and the identity-disjoint two-fold split described above was adopted. Therefore, the values reported in this paper are not directly comparable with published SYSU-MM01 results, and all models compared in Section 4.3.3 were retrained and evaluated under the same protocol. In addition, because one query image is used per identity, a single query corresponds to $1/247 \approx 0.40$ percentage points (pp) of Rank-1 accuracy on SYSU-MM01, $1/206 \approx 0.49$ pp on DBPerson-Recog-DB1, and approximately 1.2 pp on the NIR subset of RGBNT201. Differences in Rank-k accuracy of this magnitude correspond to one query and should be interpreted with caution.

**Table 1.** Configuration of images for each fold in the dataset used in our experiment.

| **Database** | | | **DBPerson-Recog-DB1** | **SYSU-MM01** | **RGBNT201** |
|---|---|---|---|---|---|
| Train Set | 1st fold | | 206 identities & 1648 images | 247 identities & 6403 images | 85 identities & 1505 images |
| | 2nd fold | | 206 identities & 1648 images | 247 identities & 6385 images | 86 identities & 1617 images |
| Validation Set | 1st fold | | 206 identities & 412 images | 247 identities & 1482 images | 85 identities & 399 images |
| | 2nd fold | | 206 identities & 412 images | 247 identities & 1482 images | 86 identities & 430 images |
| Test Set | 1st fold | Query | 206 identities & 206 images | 247 identities & 247 images | 86 identities & 86 images |
| | | Gallery | 206 identities & 412 images | 247 identities & 494 images | 86 identities & 172 images |
| | 2nd fold | Query | 206 identities & 206 images | 247 identities & 247 images | 85 identities & 85 images |
| | | Gallery | 206 identities & 412 images | 247 identities & 494 images | 85 identities & 170 images |

### *4.2. Training and Validation*

In this study, experiments were conducted with the same settings for DBPerson-Recog-DB1, SYSU-MM01, and RGBNT201. The input images were resized to a fixed resolution (384 × 128), and then random horizontal flip and random erasing were applied. For the batch configuration, 6 identities were selected per batch, and 2 images were used for each identity, resulting in a total batch size of 12. The number of epochs was set to 300, Adam was used as the optimizer, the initial learning rate was set to $6 \times 10^{-4}$, and weight decay was set to $5 \times 10^{-4}$. For the scheduler of the learning rate, a cosine annealing method including warm-up was applied. During the initial 10 epochs, warm-up was performed by linearly increasing the learning rate, and then cosine decay was applied until the final epoch to gradually decrease it.

During QAT, both the weights and activations of Extreme LightMBN were simulated at 8-bit precision using fake quantization, and the trained model was converted to INT8 for inference. The remaining hyperparameters were set as follows: the label-smoothing coefficient $\varepsilon = 0.1$, the KD temperature $T = 1.5$, and the KLW scaling factor $\delta = 0.01$. For the MS loss [38], the margin was 0.8, the similarity threshold was 0.5, and the scaling factors for positive and negative pairs were 2 and 40, respectively. $T$, $\delta$, and the MS margin were fixed in preliminary experiments conducted under the same protocol (i.e., also evaluated on the test set) and were not tuned further. For all models, including the compared models, the test set was evaluated periodically during training, and the checkpoint with the highest Rank-1 accuracy (with mAP as a tie-breaker) is reported in Section 4.3. Because both the checkpoint and the loss weights were selected using test-set results, the reported values may be optimistically biased; the validation set described in Section 4.1 contains images of the training identities and was therefore not used for model selection. For reference, when the final (300th-epoch) checkpoint was used instead, the proposed model obtained an mAP of 0.7583 and a Rank-1 accuracy of 0.7773 on SYSU-MM01, compared with 0.6752 and 0.6943 for OSNet, so the proposed model still outperformed OSNet by a similar margin. More generally, at the final recorded epoch, the proposed model achieved the highest mAP and Rank-1 accuracy among all compared models whose training logs were available (10 models

on SYSU-MM01 and 8 models on DBPerson-Recog-DB1); on DBPerson-Recog-DB1, its final-epoch mAP and Rank-1 accuracy were 0.9861 and 0.9854, compared with 0.9775 and 0.9782 for MobileNet V4. These final-epoch comparisons support a similar ranking among the models with available training logs, so the ranking is not an artifact of the test-based checkpoint selection; however, they do not eliminate the potential bias arising from the test-informed selection of the loss weights and other hyperparameters. The final-epoch values above are from the original training logs, whereas the three-seed results in Section 5.1 and the quantization comparison below were obtained by re-evaluating the final QAT checkpoints with the re-evaluation procedure described in Section 5.1. QAT was implemented with the eager-mode quantization API of PyTorch (torch.ao.quantization) using the default QAT configuration for the QNNPACK backend: weights were fake-quantized to signed 8-bit integers with per-tensor symmetric quantization, and activations were fake-quantized to unsigned 8-bit integers with per-tensor affine quantization, both using moving-average min–max observers. Convolution, batch normalization, and ReLU layers were not fused, so batch normalization layers were quantized separately. A quantization stub was placed at the input, and dequantization was applied to the outputs of each BNNeck; the KLW layer was excluded from quantization and is not executed at inference. After training, the model was converted to an INT8 model with torch.ao.quantization.convert. All INT8 student results reported in Section 4.3 were obtained by evaluating this converted INT8 model on the CPU, not by fake-quantized simulation, whereas the teacher, the FP32 student, and all compared models were evaluated in FP32. To quantify the effect of quantization, the final QAT weights of the students retrained with three random seeds (Section 5.1) were evaluated with fake quantization disabled (FP32), with fake quantization, and after conversion to INT8, using the same re-evaluation procedure as the query-level analysis in Section 5.1. Averaged over the three seeds and two folds, the mAP was 0.7557, 0.7582, and 0.7534, and the Rank-1 accuracy was 0.7632, 0.7638, and 0.7625, respectively. The FP32-to-INT8 mAP gap was 0.0023 on average and at most 0.0049, indicating that the converted INT8 model retains the accuracy of the FP32 weights learned with QAT. During the per-epoch validation, the QAT student was evaluated without disabling its observers, so the validation images, which belong to the training identities, also updated the activation ranges; no test image was used for this purpose. The observers were disabled in the re-evaluation described above, which prevents further updates but does not reverse the updates accumulated during training; the effect of these validation-based updates on the final model was not isolated. The validation loader used the training transform and random sampling, and validation was used only to monitor the curves in Figure 8. The teacher and student models were trained for 300 epochs on a single GPU. Excluding the per-epoch test evaluation, the training time per fold on SYSU-MM01 was approximately 19 h for the teacher and 8 h for the student, including the forward pass of the frozen teacher during distillation. The backbones were initialized with the ImageNet-pretrained OSNet weights released by the authors of OSNet [21] ($\beta = 1.0$ for the teacher and $\beta = 0.5$ for the student).

Meanwhile, Figure 8 shows the changes in training and validation accuracy and loss of the proposed student on SYSU-MM01 during the training process; the accuracy is the identity classification accuracy obtained from the average of the logits of the eight student classifiers. As the epoch increases, the training loss tends to decrease and the training accuracy tends to increase, which indicates that the proposed model is converging stably. The validation loss also decreases and the validation accuracy increases. However, because the validation set described in Section 4.1 consists of held-out images of the training identities, these curves measure closed-set classification and cannot by themselves show that the model generalizes to unseen identities or does not overfit. To examine retrieval behavior on unseen identities, the mAP and Rank-1 accuracy recorded on the identity-disjoint test

set of SYSU-MM01 during training were analyzed. Averaged over the two folds, the mAP remained stable during the last 100 epochs (0.7537 ± 0.0038, range 0.7438–0.7635, epochs 200–300) with a slightly positive linear trend (+0.0054 per 100 epochs), and the Rank-1 accuracy was 0.7725 ± 0.0090. Therefore, no decline in retrieval performance on unseen identities, which would indicate overfitting, was observed in the late stage of training. These epoch-averaged values are approximately 0.011 lower in mAP than the best-epoch value reported in Section 4.3, which gives a rough indication of the optimistic bias of the test-based checkpoint selection described in Section 4.2, although this difference may also include other training effects. Evaluating mAP and Rank-1 on an identity-disjoint validation set constructed from the training identities remains future work.

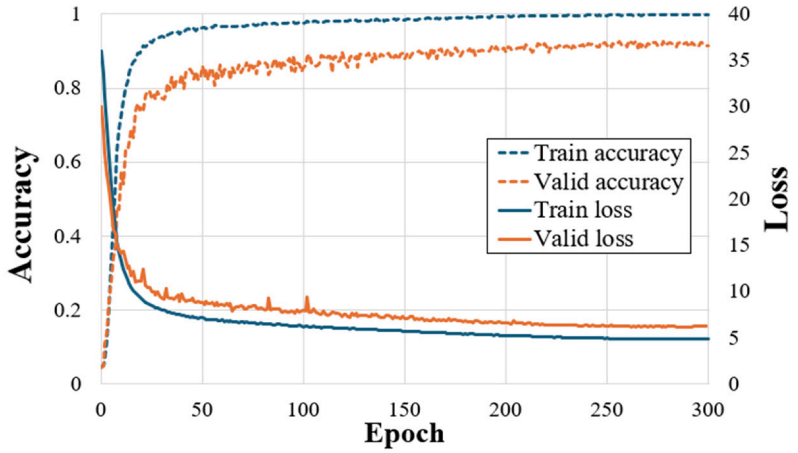



**Figure 8.** Training and closed-set validation classification accuracy and loss curves. The validation set consists of held-out images of the training identities; therefore, these curves reflect closed-set classification rather than retrieval generalization to unseen identities.

### *4.3. Test Results*

#### 4.3.1. Evaluation Metrics

In this study, mean average precision (mAP) and cumulative matching characteristic (CMC)-based Rank-1, Rank-5, and Rank-10 accuracy were used to evaluate Re-ID performance [1]. Rank-k accuracy refers to the ratio at which an image of the same person is included within the top k ranks when gallery images are sorted in order of similarity for each query image. Therefore, Rank-1 evaluates the accuracy of the top-ranked retrieval result, and Rank-5 and Rank-10 evaluate whether the correct image is included within the top 5 and 10 retrieval results, respectively. Meanwhile, mAP is a metric that comprehensively evaluates not only whether the correct answer exists at a single rank but also how highly the correct images are placed in the entire retrieval ranking. It is defined as the average value of average precision (AP) calculated for each query. For all models, the features of each image and of its horizontally flipped version were summed, re-ranking was not used, and the results of the two folds were averaged with equal weight. For the proposed model, each of the eight BNNeck descriptors was L2-normalized and concatenated into a 2048-dimensional descriptor, and the distance was computed as one minus the inner product of the concatenated descriptors, which ranks the gallery in the same order as the cosine distance. The mAP used in this study is expressed as Equation (15), where $Q$ denotes the total number of queries.

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q) \tag{15}$$

#### 4.3.2. Ablation Study

As the first ablation study, the parameter counts and computational complexity of LightMBN and Extreme LightMBN were compared using an input resolution of 384 × 128. As shown in Table 2, Extreme LightMBN reduces the parameter count from 7.54 M to 2.35 M and the computational cost from 2.42 to 0.66 GFLOPs while preserving a multi-branch structure.

**Table 2.** Comparison of parameter counts and computational complexity between LightMBN and Extreme LightMBN.

| Model | Parameters (M) | FLOPs (G) |
|---|---|---|
| LightMBN | 7.54 | 2.42 |
| Extreme LightMBN | 2.35 | 0.66 |

In the second ablation study, the main model configurations were compared on SYSU-MM01. Table 3 shows that Enhanced LightMBN outperforms the original LightMBN, indicating the benefit of the teacher-side enhancements. EQT-KD improves the FP32 student's mAP from 0.7449 to 0.7646 and Rank-1 accuracy from 0.7733 to 0.8037. The EQT-KD student also exceeds the original LightMBN in these two metrics, although a performance gap remains relative to Enhanced LightMBN. Nevertheless, this gap is small (0.0120 in mAP and 0.0101 in Rank-1 accuracy), even though the student has far fewer parameters and runs in INT8 precision; this indicates that most of the retrieval capability of the teacher is transferred through EQT-KD.

**Table 3.** Performance comparison of the main model configurations on SYSU-MM01.

| Model | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| LightMBN [35] | 0.7637 | 0.7915 | 0.9352 | 0.9859 |
| Enhanced LightMBN (teacher) | 0.7766 | 0.8138 | 0.9494 | 0.9899 |
| Extreme LightMBN (FP32, without KD) | 0.7449 | 0.7733 | 0.9352 | 0.9697 |
| Extreme LightMBN (EQT-KD) | 0.7646 | 0.8037 | 0.9373 | 0.9839 |

As the third ablation study, an experiment was conducted to investigate the performance improvement effect of the URT block, WSC structure, and PWG module in the teacher model. Table 4 shows that adding the URT block, WSC structure, or PWG module individually improved Re-ID performance over the baseline, and that the highest performance was recorded when all three components were applied. However, the individual gains are small, and the combinations are not additive; for example, URT and WSC together performed worse than either component alone. These differences should therefore be interpreted with caution in view of the query-level uncertainty discussed in Section 5.1. In all configurations without PWG, the part branch uses the same three-part split (head, upper body, and lower body), so the rows with and without PWG isolate the effect of the gate from that of the split.

In the fourth ablation study, the effect of various KD methods on the performance of the student model was analyzed. Table 5 compares the performance by applying L1-, L2-, and KL divergence-based KD from the teacher model to the student model in the QAT environment, and analyzes the difference from the proposed EQT-KD method. It was confirmed that KL divergence-based KD generally shows better performance than L1- or L2-based KD. This can be interpreted as KL divergence helping the feature representation to be learned stably by more effectively aligning the logit distributions of the teacher and student models. In addition, when the proposed EQT-KD method was applied, the highest mAP and

Rank-1 accuracy were recorded. This shows that the representation capability of the student model can be effectively improved by learning while considering the information loss that occurs during the FQ process. These results show that the proposed KD strategy can effectively alleviate the problem of performance degradation that occurs in the quantization environment. For reference, when the same KD was applied to the FP32 student without QAT, the mAP and Rank-1 accuracy were 0.7609 and 0.7936 on SYSU-MM01 and 0.9858 and 0.9879 on DBPerson-Recog-DB1, respectively; thus, the INT8 student trained with EQT-KD (0.7646 and 0.8037 on SYSU-MM01 and 0.9883 and 0.9903 on DBPerson-Recog-DB1) reached the accuracy of the FP32 student trained with KD. Relative to uniform KL-KD, however, the additional gain of EQT-KD is small (paired query-level mAP difference of +0.0059, 95% CI: [−0.0094, +0.0212]) and within the query-level uncertainty, and the Rank-5 and Rank-10 accuracies were slightly lower; most of the improvement over the student without KD is therefore attributable to distillation combined with QAT.

**Table 4.** Performance comparison according to the presence of each component in the teacher model. In all configurations without PWG, the part branch uses the three-part split (head, upper body, and lower body).

| URT | WSC | PWG | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|---|---|
| | | | 0.7637 | 0.7915 | 0.9352 | 0.9859 |
| ✓ | | | 0.7683 | 0.7996 | 0.9413 | 0.9778 |
| | ✓ | | 0.7684 | 0.7976 | 0.9454 | 0.9859 |
| | | ✓ | 0.7722 | 0.7976 | 0.9474 | 0.9778 |
| ✓ | ✓ | | 0.7656 | 0.7895 | 0.9454 | 0.9859 |
| ✓ | | ✓ | 0.7764 | 0.8118 | 0.9434 | 0.9859 |
| | ✓ | ✓ | 0.7705 | 0.7996 | 0.9494 | 0.9859 |
| ✓ | ✓ | ✓ | 0.7766 | 0.8138 | 0.9494 | 0.9899 |

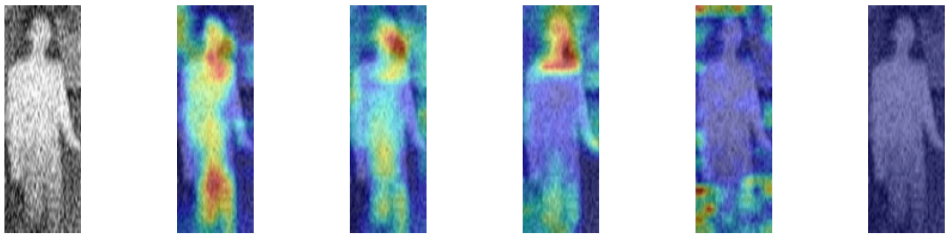
**Table 5.** Comparison of KD methods for the student model.

| Method | | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|---|
| QAT | L1-KD | 0.7546 | 0.7875 | 0.9373 | 0.9798 |
| | L2-KD | 0.7556 | 0.7915 | 0.9454 | 0.9778 |
| | KL-KD | 0.7620 | 0.7996 | 0.9474 | 0.9940 |
| EQT-KD (proposed) | | 0.7646 | 0.8037 | 0.9373 | 0.9839 |

In the fifth ablation study, Extreme LightMBN was evaluated in FP32 and using QAT at 8-, 6-, 4-, and 2-bit precision without knowledge distillation under the same training and evaluation settings. Table 6 shows that reducing the precision from 8 to 6 bits decreases mAP from 0.7345 to 0.6707, while 4- and 2-bit quantization cause substantially larger reductions. These results characterize the quantization sensitivity of the student trained without KD. For 6, 4, and 2 bits, the default QNNPACK QAT configuration was modified to restrict the integer ranges of the weights to $[-2^{b-1}, 2^{b-1} - 1]$ and those of the activations to $[0, 2^{b} - 1]$ for b bits, and the models were converted and evaluated with 8-bit integer storage; therefore, these results simulate low-bit quantization and do not represent packed low-bit inference. Comparing Tables 5 and 6, EQT-KD improves the 8-bit student's mAP from 0.7345 to 0.7646 and Rank-1 accuracy from 0.7652 to 0.8037. Figure 9 presents an example of the corresponding Grad-CAM visualizations at different precisions.

**Table 6.** Performance of Extreme LightMBN without knowledge distillation at different quantization bit-widths on SYSU-MM01.

| Precision | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| FP32 | 0.7449 | 0.7733 | 0.9352 | 0.9697 |
| QAT-INT8 | 0.7345 | 0.7652 | 0.9292 | 0.9778 |
| QAT-INT6 | 0.6707 | 0.7247 | 0.9109 | 0.9595 |
| QAT-INT4 | 0.2327 | 0.2267 | 0.4271 | 0.5507 |
| QAT-INT2 | 0.0157 | 0.0040 | 0.0202 | 0.0405 |



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

**Figure 9.** Original SYSU-MM01 image and Grad-CAM visualizations of Extreme LightMBN trained without knowledge distillation at different precisions: (**a**) original input image, (**b**) FP32, (**c**) QAT-INT8, (**d**) QAT-INT6, (**e**) QAT-INT4, and (**f**) QAT-INT2. In the Grad-CAM maps, red and yellow indicate regions with a larger contribution to the predicted identity and blue indicates a smaller contribution; each map is min–max normalized, and the examples are qualitative.

In the sixth ablation study, the effect of each loss weight in the proposed total loss function on performance was analyzed. From Table 7, when $\lambda_{ce} = 0.5$ was fixed, the setting with $\lambda_{ms} = 0.5$ and $\lambda_{kd} = 0.25$ achieved the highest Rank-1 accuracy. $\lambda_{ce}$ of 0.5 was fixed to maintain the same identity classification criterion across all settings. This setting effectively maintained the balance between strengthening the discriminative capability of the embedding feature and transferring teacher knowledge. In contrast, when $\lambda_{ms}$ was low or $\lambda_{kd}$ was excessively large, performance decreased. Therefore, in this study, the final loss weights were set to $\lambda_{ce} = 0.5$, $\lambda_{ms} = 0.5$, and $\lambda_{kd} = 0.25$. It should be noted that Table 7 is a sensitivity analysis on the test set rather than a selection on an independent validation set. The setting used for the final model ($\lambda_{ms} = 0.5$ and $\lambda_{kd} = 0.25$) achieved the highest Rank-1 accuracy but not the highest mAP (0.7687 for $\lambda_{ms} = 0.25$ and $\lambda_{kd} = 0.5$), and its Rank-1 advantage over the other settings (at most 2.2 pp) is within the query-level uncertainty reported in Section 5.1. Nevertheless, in all four settings, the mAP of the proposed student (0.7539–0.7687) was higher than that of every compared model in Table 8, indicating that the main conclusion does not depend on the particular choice of loss weights. $\lambda_{ce}$ was fixed at 0.5 and was not varied, and selecting all loss weights on an identity-disjoint validation set remains a limitation of this study.

**Table 7.** Sensitivity analysis of the loss weights on the test set of SYSU-MM01.

| $\lambda_{ce}$ | $\lambda_{ms}$ | $\lambda_{kd}$ | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|---|---|
| 0.5 | 0.25 | 0.25 | 0.7539 | 0.7814 | 0.9434 | 0.9839 |
| 0.5 | 0.25 | 0.5 | 0.7687 | 0.7915 | 0.9434 | 0.9859 |
| 0.5 | 0.5 | 0.25 | 0.7646 | 0.8037 | 0.9373 | 0.9839 |
| 0.5 | 0.5 | 0.5 | 0.7658 | 0.7976 | 0.9434 | 0.9798 |

**Table 8.** Comparison with state-of-the-art models on the SYSU-MM01 dataset.

| Model | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| ResNet50 [36] | 0.6389 | 0.6761 | 0.8846 | 0.9434 |
| PCB [44] | 0.6347 | 0.6782 | 0.8684 | 0.9332 |
| OSNet [21] | 0.6878 | 0.7328 | 0.9069 | 0.9555 |
| TransReID [45] | 0.4403 | 0.4555 | 0.7207 | 0.8361 |
| MaxViT [46] | 0.6052 | 0.6316 | 0.8523 | 0.9150 |
| Swin Transformer V2 [47] | 0.6409 | 0.6498 | 0.8725 | 0.9393 |
| OADE-Net [7] | 0.3221 | 0.4049 | 0.6864 | 0.7935 |
| NFormer [48] | 0.6446 | 0.6640 | 0.8907 | 0.9352 |
| ConvNeXt V2 [49] | 0.5455 | 0.5567 | 0.8037 | 0.8846 |
| StarNet [50] | 0.6374 | 0.6721 | 0.9089 | 0.9413 |
| MobileNet V4 [51] | 0.6538 | 0.6680 | 0.8988 | 0.9555 |
| RapidNet [52] | 0.5635 | 0.5871 | 0.8178 | 0.8988 |
| MambaVision [53] | 0.6618 | 0.7025 | 0.9028 | 0.9494 |
| WSE-Net [31] | 0.5577 | 0.5830 | 0.8562 | 0.9291 |
| KnightReID [30] | 0.5877 | 0.6336 | 0.8219 | 0.9029 |
| Proposed model | 0.7646 | 0.8037 | 0.9373 | 0.9839 |

#### 4.3.3. Comparison of Proposed Model and the State-of-the-Art (SOTA) Models

To evaluate the performance of the proposed model, Re-ID performance was compared with various SOTA models on SYSU-MM01, DBPerson-Recog-DB1, and the NIR subset of RGBNT201. All models were trained and evaluated using the same dataset splits, query–gallery protocol, and evaluation metrics, and with the same training recipe: Adam (learning rate $6 \times 10^{-4}$, weight decay $5 \times 10^{-4}$) with cosine annealing for 300 epochs, batches of 6 identities × 2 images, cross-entropy loss with label smoothing and MS loss (weights of 0.5 each, margin of 0.8), and random erasing. All backbones were initialized with ImageNet-pretrained weights except RapidNet, which was trained from scratch. The input size was 384 × 128 pixels except for MaxViT (224 × 224) and Swin Transformer V2 (256 × 256), whose architectures require these input sizes. Because large models trained on a few thousand images with this batch size may underperform for reasons unrelated to their architecture, their results should be interpreted as performance under this low-data setting rather than as a comparison of the architectures themselves. These comparisons follow the single-modality IR protocol described in Section 4.1.

From the results of Table 8, the proposed model recorded the highest mAP and Rank-1 accuracy among the compared models. Compared with the FP32 LightMBN in Table 3, the INT8 student achieves a comparable mAP (0.7646 vs. 0.7637) and a higher Rank-1 accuracy (0.8037 vs. 0.7915) with 31% of its parameters (2.35 M vs. 7.54 M), which indicates comparable accuracy at a much lower cost rather than a large gain in accuracy. This experiment was conducted in the low-resolution IR SYSU-MM01 environment, where color and texture information are limited and the effects of noise and blur are large. Under these conditions, CNN-based models showed relatively stable performance compared with transformer-series models; a possible interpretation is that local feature extraction is better suited to these images than patch-level modeling, although pretraining and the common training recipe may also contribute. OSNet recorded the second-highest performance, and the remaining gap may be related to the absence of components that address part-level and

noise-related degradation, although this was not tested directly. Similarly, OADE-Net was designed for feature enhancement of HR IR images, which may limit its performance in the low-resolution environment of this experiment, although its optimization under the common recipe may also contribute.

From the results of Table 9, the proposed model achieved the highest mAP and tied with MobileNet V4 for the highest Rank-1 accuracy of 0.9903. In DBPerson-Recog-DB1, overall high performance was observed, which is considered to be because the distinction between identities is relatively clearer than in SYSU-MM01. Models such as ConvNeXt V2 and RapidNet showed limited performance because they could not sufficiently compensate for the effects of texture deficiency, blur, and noise that occur in low-resolution IR images. Because the results on this dataset are close to saturation, the differences among the stronger models are small.

**Table 9.** Comparison with state-of-the-art models on the DBPerson-Recog-DB1 dataset.

| Model | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| ResNet50 [36] | 0.9775 | 0.9854 | 1.0000 | 1.0000 |
| PCB [44] | 0.9748 | 0.9830 | 1.0000 | 1.0000 |
| OSNet [21] | 0.9811 | 0.9830 | 0.9952 | 0.9976 |
| TransReID [45] | 0.8273 | 0.8398 | 0.9588 | 0.9806 |
| MaxViT [46] | 0.9619 | 0.9757 | 0.9879 | 0.9951 |
| Swin Transformer V2 [47] | 0.9745 | 0.9830 | 0.9927 | 0.9952 |
| OADE-Net [7] | 0.9410 | 0.9757 | 0.9951 | 0.9976 |
| NFormer [48] | 0.9605 | 0.9636 | 0.9952 | 0.9952 |
| ConvNeXt V2 [49] | 0.9010 | 0.9199 | 0.9685 | 0.9806 |
| StarNet [50] | 0.9744 | 0.9854 | 0.9903 | 0.9952 |
| MobileNet V4 [51] | 0.9804 | 0.9903 | 0.9952 | 1.0000 |
| RapidNet [52] | 0.8318 | 0.8374 | 0.9588 | 0.9733 |
| MambaVision [53] | 0.9735 | 0.9830 | 0.9903 | 0.9976 |
| WSE-Net [31] | 0.9001 | 0.9709 | 0.9879 | 0.9927 |
| KnightReID [30] | 0.9740 | 0.9757 | 0.9927 | 0.9952 |
| Proposed model | 0.9883 | 0.9903 | 0.9951 | 0.9976 |

On the NIR subset of RGBNT201, the proposed model achieved the highest mAP and Rank-1 accuracy among the compared models, with values of 0.7473 and 0.7547, respectively (Table 10). Compared with OSNet, which obtained the second-highest values in these two metrics, the improvements were 2.58 and 1.18 percentage points, respectively. However, the proposed model did not achieve the highest Rank-5 or Rank-10 accuracy, and this finding is limited to the NIR subset of RGBNT201 evaluated under the protocol of this study.

**Table 10.** Comparison with state-of-the-art models on the NIR subset of RGBNT201 under the evaluation protocol used in this study.

| Model | mAP | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| ResNet50 [36] | 0.6800 | 0.7079 | 0.7898 | 0.8481 |
| PCB [44] | 0.6886 | 0.7197 | 0.7783 | 0.8545 |
| OSNet [21] | 0.7215 | 0.7429 | 0.8425 | 0.8660 |
| TransReID [45] | 0.6491 | 0.6961 | 0.7664 | 0.8015 |
| MaxViT [46] | 0.6624 | 0.7138 | 0.7838 | 0.8247 |
| Swin Transformer V2 [47] | 0.6791 | 0.7077 | 0.8130 | 0.8423 |
| OADE-Net [7] | 0.6427 | 0.6612 | 0.7955 | 0.8891 |
| NFormer [48] | 0.6132 | 0.6198 | 0.7605 | 0.8542 |
| ConvNeXt V2 [49] | 0.6414 | 0.6787 | 0.7607 | 0.8015 |
| StarNet [50] | 0.6926 | 0.7255 | 0.7899 | 0.8600 |
| MobileNet V4 [51] | 0.7010 | 0.7313 | 0.8074 | 0.8249 |
| RapidNet [52] | 0.6294 | 0.6552 | 0.7370 | 0.8131 |
| MambaVision [53] | 0.7131 | 0.7139 | 0.7956 | 0.8484 |
| WSE-Net [31] | 0.5615 | 0.5966 | 0.7077 | 0.7722 |
| KnightReID [30] | 0.6442 | 0.6846 | 0.7197 | 0.7608 |
| Proposed model | 0.7473 | 0.7547 | 0.8190 | 0.8716 |

## 5. Discussions

### *5.1. Statistical Analysis*

To examine whether the performance difference between the proposed model and the second-best model (OSNet) on SYSU-MM01 is statistically meaningful, two complementary analyses were conducted, and the results are shown in Figure 10. First, at the fold level, a two-tailed independent-samples *t*-test [54] and Cohen's d [55] were computed from the mAP values of the two folds (n = 2 per model). This test yielded t = 4.95 with 2 degrees of freedom ($p = 0.0385$) and d = 4.95. This test assumes normally distributed fold-level values with equal variances, which cannot be verified with two observations per model; therefore, this fold-level test is reported only as an exploratory reference, and the inferential conclusions rely on the paired query-level analyses described below. However, because only two observations are available per model, this test has very limited statistical power, and the fold-level standard deviations do not support conclusions about the stability of either model. Second, to quantify the uncertainty arising from the limited number of query images, a query-level analysis was performed using all 494 test queries of the two folds (247 per fold). The 95% confidence intervals (CIs) were obtained from 10,000 paired bootstrap resamples of the queries, and a Wilcoxon signed-rank test was applied to the paired per-query average precision values. For this analysis, the stored checkpoints (INT8 for the proposed model and FP32 for OSNet) were re-evaluated on our desktop computer; owing to differences in the inference runtime and environment, the re-evaluated values (mAP of 0.7649 and Rank-1 accuracy of 0.7976 for the proposed model, and 0.6871 and 0.7328 for OSNet) differ slightly from those in Table 8. As summarized in Table 11, the mAP difference was +0.0779 (95% CI: [+0.0581, +0.0980]) and the Rank-1 difference was +0.0649 (95% CI: [+0.0283, +0.1012]). Both intervals exclude zero, the Wilcoxon signed-rank test yielded $p = 1.03 \times 10^{-12}$, and the mAP difference was positive in each fold (+0.0683 [+0.0406, +0.0965] and +0.0871 [+0.0588, +0.1164] in the first and second folds, respectively). These results indicate that the

improvement over OSNet exceeds the uncertainty caused by query sampling. In contrast, the 95% CI of Rank-1 accuracy spans approximately ±3.5 pp, so differences smaller than about 1 pp, such as those among several configurations in Tables 4 and 5, and 7, are within this uncertainty and should be interpreted with caution. Furthermore, the query-level analysis does not capture variability due to training randomness or the choice of identity split. To assess the variability due to training randomness, the proposed model was retrained three times on both folds of SYSU-MM01 with different random seeds (251, 1, and 2), and the final-epoch INT8 model of each run was re-evaluated with the same procedure as the query-level analysis. Over the six runs, the mAP was 0.7534 ± 0.0048 (mean ± standard deviation; range 0.7478–0.7612), and the Rank-1 accuracy was 0.7625 ± 0.0055 (range 0.7530–0.7692); averaged over the two folds, the mAP for seeds 251, 1, and 2 was 0.7557, 0.7553, and 0.7493, respectively. Even the lowest values exceeded the final-epoch mAP and Rank-1 accuracy of OSNet (0.6752 and 0.6943; Section 4.2). Evaluating the variability due to the choice of identity split with additional splits remains a limitation of this study.

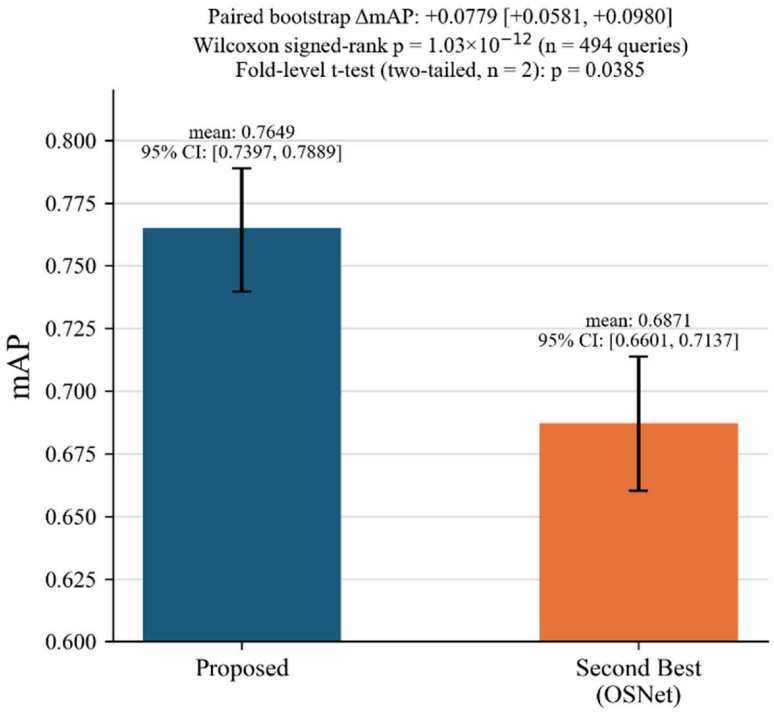



**Figure 10.** Statistical analysis of mAP between the proposed model and the second-best model (OSNet) on SYSU-MM01. Bars show the mean over 494 test queries of the two folds, and error bars show 95% confidence intervals from 10,000 paired bootstrap resamples. The paired mAP difference, the Wilcoxon signed-rank test on per-query average precision, and the fold-level two-tailed *t*-test (n = 2 per model) are shown above the plot.

The same query-level analysis was extended to the other compared models and to DBPerson-Recog-DB1 by re-evaluating the stored checkpoints of ResNet50, OSNet, TransReID, MaxViT, Swin Transformer V2, ConvNeXt V2, StarNet, MobileNet V4, and RapidNet on SYSU-MM01 and of ResNet50, TransReID, MaxViT, ConvNeXt V2, StarNet, and MobileNet V4 on DBPerson-Recog-DB1. On SYSU-MM01, the paired 95% CIs of both the mAP and Rank-1 differences between the proposed model and each of the nine re-evaluated models excluded zero; apart from OSNet, the smallest differences were obtained for MobileNet V4 (mAP: +0.1114 [+0.0895, +0.1339]; Rank-1: +0.1297 [+0.0911, +0.1680]). On DBPerson-Recog-DB1, where the accuracies are close to saturation, the paired CIs of the mAP differences also excluded zero for all six re-evaluated models, including MobileNet

V4 (+0.0073 [+0.0019, +0.0135]), whereas the Rank-1 accuracy of the proposed model was tied with that of MobileNet V4 (0.9903). These results show that the mAP advantage of the proposed model is consistent across the compared models and both datasets and exceeds the uncertainty caused by the limited number of queries.

**Table 11.** Query-level comparison between the proposed model and the second-best model (OSNet) on SYSU-MM01. The values are the re-evaluated mAP and Rank-1 accuracy over the 494 test queries of the two folds with 95% CIs from 10,000 paired bootstrap resamples; the difference is given as the proposed model minus OSNet. The Wilcoxon signed-rank test on the per-query average precision yielded $p = 1.03 \times 10^{-12}$.

| Model | mAP [95% CI] | Rank-1 [95% CI] |
|---|---|---|
| Proposed model | 0.7649 [0.7397, 0.7889] | 0.7976 [0.7631, 0.8320] |
| OSNet | 0.6871 [0.6601, 0.7137] | 0.7328 [0.6923, 0.7713] |
| Difference | +0.0779 [+0.0581, +0.0980] | +0.0649 [+0.0283, +0.1012] |

### *5.2. Analysis of Extracted Features by Gradient-Weighted Class Activation Mapping (Grad-CAM)*

To analyze which regions in the input image the proposed model focuses on when performing person re-identification, Grad-CAM [56–58] was applied. Grad-CAM was extracted from the last convolution block in the global branch of Extreme LightMBN. Figure 11 presents the Grad-CAM visualization results on DBPerson-Recog-DB1 and SYSU-MM01. The visualization results showed that the model tended to focus more on human body regions than on the background. In both the DBPerson-Recog-DB1 and SYSU-MM01 datasets, activation was observed in regions important for person identification, such as the upper body, torso, and lower body. These examples suggest that the proposed model relies mainly on cues from the person rather than from the background. However, activation outside the body region is also visible in some maps, and these few examples do not by themselves establish robustness to low resolution and noise. The Grad-CAM maps were computed for the QAT student model before INT8 conversion, using the output of the last block of the global branch as the target layer and the logit of the predicted class of the main global classifier as the target score. The channel weights were obtained by spatially averaging the gradients, and the weighted activation map was passed through ReLU, upsampled to 384 × 128 pixels with bilinear interpolation, min–max normalized, and overlaid on the input image with equal weights. These visualizations are qualitative examples and are not intended as a quantitative evaluation of the attended regions.

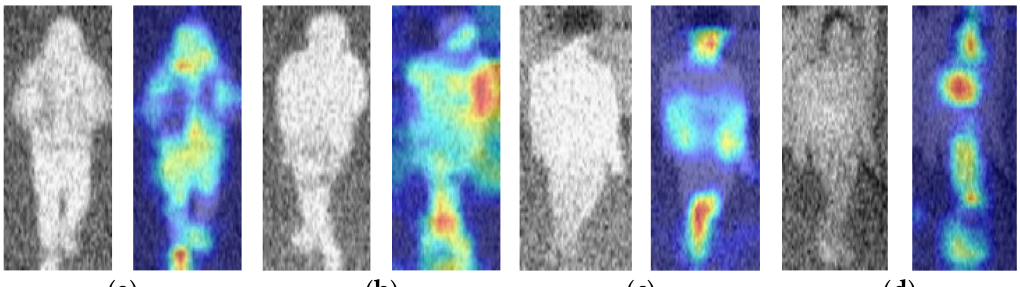



**Figure 11.** Grad-CAM results on DBPerson-Recog-DB1 and SYSU-MM01. In each pair, the left image is the original IR image and the right image is the corresponding Grad-CAM map. (**a**,**b**) are from DBPerson-Recog-DB1, and (**c**,**d**) are from SYSU-MM01. Red and yellow indicate regions with a larger contribution to the predicted identity and blue indicates a smaller contribution; each map is min–max normalized, and the examples are qualitative.

### *5.3. Analysis of Genuine and Impostor Matching Cases*

To analyze the matching characteristics of the proposed model, examples of genuine pairs and impostor pairs were visually compared. In Figure 12, the genuine pairs showed results in which the same person was stably matched while overcoming pose variations and differences between front and back views, even in an infrared environment with low resolution and noise. In contrast, the impostor pairs showed misrecognition results due to similar appearance, clothing, and intensity distribution, even though they were different persons, in a situation where detailed appearance information was limited by low resolution and noise. This suggests that difficult misrecognition cases still exist under poor image conditions. Figure 12 shows a small number of examples for illustration and does not represent the overall distribution of matching errors.

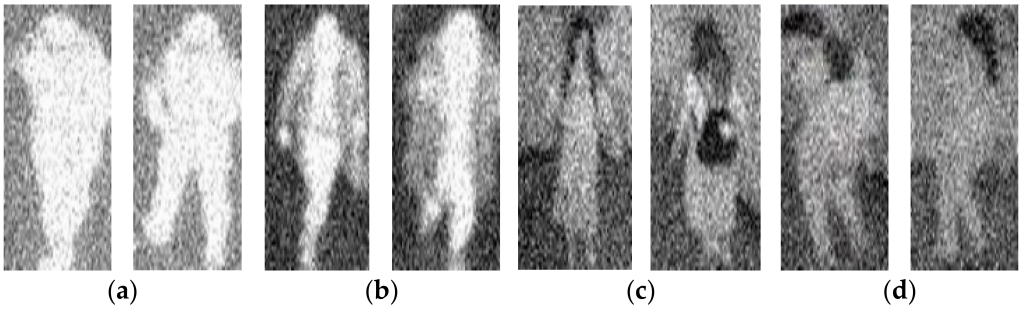



**Figure 12.** Genuine and impostor matching examples from DBPerson-Recog-DB1 and SYSU-MM01. (**a**,**c**) denote genuine pairs, while (**b**,**d**) denote impostor pairs; (**a**,**b**) are from DBPerson-Recog-DB1 and (**c**,**d**) are from SYSU-MM01. The examples are qualitative illustrations and do not represent the overall distribution of matching results.

### *5.4. Computational Cost*

The computational cost of the proposed model was evaluated based on the number of model parameters, GPU memory usage, and floating-point operations (FLOPs). Table 12 shows the comparison results of the number of parameters, GPU memory usage, FLOPs, and power consumption for various models. The proposed model has low computational complexity with 2.35 M parameters and 0.66 GFLOPs; the parameter count includes the classifier heads used only during training (Section 3.3.1), and FLOPs were counted as multiply–accumulate operations of the convolution and linear layers used at test time with the complexity tool of torchreid, and its GPU memory usage also showed the lowest level among the compared models at 3.57 MB. In addition, its measured power consumption was the lowest at 1.05 W; however, the differences from several models (1.09–1.14 W for StarNet, RapidNet, NFormer, and OADE-Net) are small. Power consumption is not equivalent to the energy consumed per inference, which also depends on the inference time, and therefore Table 12 does not by itself establish a difference in energy efficiency. These results indicate that the proposed model has low computational complexity and low reported memory usage, which make it suitable for resource-limited edge and mobile environments. The FLOPs and inference times in Tables 12 and 13 were measured at the input size used for each model (384 × 128 pixels, except for MaxViT and Swin Transformer V2; Section 4.3.3). Because the information content of the input corresponds to the 64 × 64 LR images, the computational cost could be reduced further with a smaller input size; this was not evaluated in this study. The GPU memory usage in Table 12 was measured with torch.cuda.memory_allocated() and therefore reflects the memory occupied by the tensors allocated for the model, which is dominated by its weights; it does not include the CUDA context, the memory reserved by the caching allocator, or framework overhead. The power consumption in Table 12 was

measured on the NVIDIA Jetson Orin Nano with the tegrastats utility during TensorRT inference, not on the desktop GPU.

**Table 12.** Comparison of computational cost, GPU memory usage, and power consumption.

| Model | Parameter (Mega) | GPU Memory Usage (MB) | FLOPs (Giga) | Power Consumption (W) |
|---|---|---|---|---|
| ResNet50 [36] | 28.21 | 103.40 | 4.01 | 1.40 |
| PCB [44] | 31.75 | 103.05 | 6.09 | 1.52 |
| OSNet [21] | 2.32 | 19.87 | 1.46 | 1.24 |
| TransReID [45] | 21.81 | 91.00 | 1.06 | 1.30 |
| MaxViT [46] | 68.83 | 306.00 | 11.30 | 1.71 |
| Swin Transformer V2 [47] | 37.94 | 438.18 | 8.57 | 1.20 |
| OADE-Net [7] | 228.92 | 1447.05 | 48.12 | 1.14 |
| NFormer [48] | 37.21 | 101.40 | 4.08 | 1.11 |
| ConvNeXt V2 [49] | 49.81 | 208.13 | 8.52 | 1.47 |
| StarNet [50] | 3.55 | 19.34 | 0.54 | 1.09 |
| MobileNet V4 [51] | 1.86 | 10.55 | 0.18 | 1.30 |
| RapidNet [52] | 18.55 | 76.08 | 1.57 | 1.10 |
| MambaVision [53] | 49.86 | 211.39 | 11.19 | 1.80 |
| WSE-Net [31] | 110.03 | 420.89 | 35.75 | 1.20 |
| KnightReID [30] | 4.85 | 20.56 | 5.18 | 1.22 |
| Proposed model | 2.35 | 3.57 | 0.66 | 1.05 |

Table 13 shows the comparison results of processing time per image (ms) and frames per second (FPS) in various hardware environments. Each value was measured under the same experimental environment, and the value in parentheses indicates FPS. The desktop GPU was measured using NVIDIA RTX 2070, Jetson was measured in the NVIDIA Jetson Orin Nano environment based on TensorRT, and Mobile NPU used ONNX inference results based on the AI Engine (Hexagon NPU) embedded in the Qualcomm Snapdragon 8 Gen 2 platform of a Samsung Galaxy Z Fold5. The proposed model recorded processing performance of 1.34 ms (744.16 FPS) on the desktop GPU, 2.90 ms (344.76 FPS) on the Jetson GPU, and 9.66 ms (103.52 FPS) on the Mobile NPU, and its latency was lower than that of most compared models. However, MobileNet V4 was faster on the desktop GPU (0.80 ms) and the Jetson GPU (1.28 ms), and StarNet was also faster on the desktop GPU (1.17 ms), whereas the proposed model had the lowest latency on the mobile NPU (9.66 ms vs. 19.90 ms for MobileNet V4). The poor mobile NPU results of the Transformer- and Mamba-based models may partly reflect operators that are not supported by the NPU and fall back to other processors, and they should not be interpreted as a comparison of the architectures themselves. Furthermore, these values measure model inference only and do not include other parts of a complete surveillance pipeline, such as image acquisition, person detection, and gallery matching; therefore, they indicate the potential for real-time operation rather than demonstrating end-to-end real-time performance. On the desktop GPU and the Jetson, the inference time was measured with the trtexec tool of TensorRT; the proposed model was built as an INT8 engine, whereas the compared models were built as FP32 engines, so part of the speed advantage of the proposed model on these platforms

comes from INT8 inference. Each model was measured with the default settings of trtexec (batch size 1, a 200 ms warm-up, and repeated inference for at least 10 iterations and 3 s). The measured times correspond to a single forward pass of one image without the flipped input, and data transfer to the device and gallery matching are not included. Retrieval accuracy after the TensorRT and NPU conversions was not evaluated; the accuracies in Tables 3–10 are those of the QNNPACK INT8 model.

**Table 13.** Comparison of inference time and FPS across different hardware platforms.

| Model | Desktop GPU | JETSON GPU | Mobile NPU |
|---|---|---|---|
| ResNet50 [36] | 2.73 (366.71) | 6.70 (149.16) | 211.00 (4.74) |
| PCB [44] | 2.78 (360.05) | 6.60 (151.53) | 198.31 (5.04) |
| OSNet [21] | 2.05 (488.04) | 6.70 (149.27) | 149.85 (6.67) |
| TransReID [45] | 3.11 (322.04) | 9.14 (109.47) | 566.07 (1.77) |
| MaxViT [46] | 9.26 (108.04) | 34.74 (28.78) | 1704.29 (0.59) |
| Swin Transformer V2 [47] | 9.54 (104.81) | 38.99 (25.65) | 2593.53 (0.39) |
| OADE-Net [7] | 33.83 (29.56) | 197.98 (5.05) | 555.31 (1.80) |
| NFormer [48] | 3.02 (330.92) | 7.69 (130.05) | 306.16 (3.27) |
| ConvNeXt V2 [49] | 6.37 (157.06) | 23.40 (42.74) | 1561.03 (0.64) |
| StarNet [50] | 1.17 (857.46) | 3.12 (320.73) | 59.22 (16.89) |
| MobileNet V4 [51] | 0.80 (1252.06) | 1.28 (782.21) | 19.90 (50.25) |
| RapidNet [52] | 3.81 (262.83) | 8.27 (120.82) | 272.13 (3.67) |
| MambaVision [53] | 5.31 (188.35) | 15.46 (64.67) | 1038.61 (0.96) |
| WSE-Net [31] | 34.72 (28.80) | 43.04 (23.23) | 396.48 (2.52) |
| KnightReID [30] | 5.24 (190.75) | 21.45 (46.63) | 215.91 (4.63) |
| Proposed model | 1.34 (744.16) | 2.90 (344.76) | 9.66 (103.52) |

For larger surveillance networks, the quantized student could extract embeddings at individual edge cameras, while a central server performs cross-camera matching. Such a deployment could reduce raw-image transmission and distribute the inference workload; the 2048-dimensional descriptor of the student requires 8 KB per image in 32-bit floating point format. Scalable gallery retrieval and adaptation to dynamic environmental and sensor changes remain topics with various applications, including face recognition [59], for future work.

## 6. Conclusions

This paper addressed lightweight IR person Re-ID under low resolution and noise by distilling an enhanced multi-branch teacher into a compact student trained with QAT. In particular, through the KLW module, which adjusts the strength of distillation based on the difference between the outputs before and after FQ, stable knowledge transfer was enabled even in the QAT environment. The main findings are as follows. First, the INT8 student achieved accuracy comparable to that of the FP32 LightMBN with 31% of its parameters, and its gap to the teacher was small, which indicates that most of the retrieval capability of the teacher can be retained after distillation and 8-bit quantization. Second, the gains of the individual teacher modules were small and not additive, so their individual contributions could not be separated reliably under the present protocol. Third, the inference time of the proposed model was lower than that of most compared models, particularly on the mobile

NPU, although MobileNet V4 was faster on the desktop and Jetson GPUs; because these measurements cover model inference only, they indicate the potential for real-time inference on edge and mobile platforms rather than end-to-end real-time operation. However, because this study was evaluated based on specific datasets, additional verification of generalization performance in various real surveillance environments is required. In particular, the LR images were synthesized from HR images with fixed degradation parameters rather than captured by real LR IR cameras, and the person images were resized to a square 64 × 64 format during the degradation and then to 384 × 128 pixels for the network input, so the aspect ratio of the person differs from that of the original images; evaluation on real LR IR surveillance data, at native resolutions, and over a range of resolutions and degradation levels remains for future work. In addition, the protocol uses one query and two gallery images per test identity without camera constraints, which is a smaller retrieval task than large-gallery cross-camera search, and the gain of KLW over uniform KL-KD was small relative to the query-level uncertainty (Section 4.3.2). In addition, because this study focuses on a single infrared modal environment, limitations exist in responding to various condition changes that occur in complex real environments. Furthermore, in the QAT environment, distribution changes caused by quantization occur, and therefore knowledge transfer from the teacher can be limited. Additional studies are required on methods for performing distillation more effectively in these situations.

The bit-width study without KD also showed that 4- and 2-bit quantization substantially reduces recognition accuracy, which indicates that extending the framework to lower bit-widths requires further study of quantization-aware distillation.

Person re-identification is a biometric surveillance technology that can affect personal privacy. The datasets used in this study were used only for academic research under the conditions set by their providers, and the deployment of the proposed model should comply with applicable privacy and data-protection regulations, including purpose limitation, data minimization, and appropriate human oversight, to prevent misuse such as unauthorized tracking of individuals.

In future work, we plan to analyze the generalization performance of the model more systematically based on datasets reflecting various real environments, further improve robustness in low-resolution and noisy IR image environments, and study distillation techniques that can more effectively transfer information of the teacher even in QAT environments, thereby advancing this research toward stable operation on actual consumer edge devices. We also plan to compare the proposed framework with QAT- and KD-equipped versions of the compared models, with lightweight Re-ID methods, and with existing QAT-KD methods; to compare KLW with fixed per-branch, non-learned, and random weighting schemes to examine scale-normalized measures of the $q_e$; and to apply EQT-KD at 6- and 4-bit precision.

**Author Contributions:** Conceptualization, H.W.L.; methodology, H.W.L.; validation, S.J.L., S.J.I., H.W.S. and I.H.J.; supervision, K.R.P.; writing—original draft preparation, H.W.L.; writing—review and editing, K.R.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25586697), and in part by the National Supercomputing Center at KISTI (KSC-2025-CRE-0485).

**Institutional Review Board Statement:** Ethical review and approval were waived for this study because only the open databases were used for our experiments and all the open databases can be freely used for academic purposes according to the instructions of the database providers.

**Informed Consent Statement:** Not applicable; this study used only open datasets. DBPerson-Recog-DB1 was collected by our research group [39], and the participants consented to the use of their images for research. According to the providers of SYSU-MM01, every pedestrian in the dataset signed a privacy license that allows the images to be used for scientific research and shown in research papers, and the dataset was obtained by submitting a signed dataset release agreement [40]. RGBNT201 was released by its authors for academic research [41] and was used only for non-commercial research in this study.

**Data Availability Statement:** This study uses the DBPerson-Recog-DB1, SYSU-MM01, and RGBNT201 datasets described in Section 4.1. The proposed source code, trained models, and experimental data configurations are available at https://github.com/coffeeboy123/EQT-KD (accessed on 15 November 2026). The datasets are available from their providers under their respective terms and are not redistributed by the authors.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References


1. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.H. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2872–2893. [CrossRef] [PubMed]
2. Liu, Y.; Li, J.; Lyu, G.; Yang, Y.; Li, Z.; Liu, Y. Hierarchical deep synergistic fusion of CNN and transformer for person re-identification. *J. King Saud Univ. Comput. Inf. Sci.* **2026**, *38*, 4. [CrossRef]
3. Li, Y.; Zhang, T.; Liu, X.; Tian, Q.; Zhang, Y.; Wu, F. Visible-infrared person re-identification with modality-specific memory network. *IEEE Trans. Image Process.* **2022**, *31*, 7165–7178. [CrossRef] [PubMed]
4. Qian, Z.; Lin, Y.; Du, B. Visible–infrared person re-identification via patch-mixed cross-modality learning. *Pattern Recognit.* **2025**, *157*, 110873. [CrossRef]
5. Lin, R.; Wang, R.; Zhang, W.; Wu, A.; Bi, Y. Joint modal alignment and feature enhancement for visible-infrared person re-identification. *Sensors* **2023**, *23*, 4988. [CrossRef] [PubMed]
6. Jung, S.Y.; Lee, D.C.; Jeong, M.S.; Jeong, S.I.; Song, H.W.; Lee, H.W.; Park, K.R. Estimation of fractal dimension and super-resolution reconstruction for person re-identification in images from infrared surveillance camera. *Fractal Fract.* **2026**, *10*, 123. [CrossRef]
7. Jeong, M.S.; Jeong, S.I.; Kang, S.J.; Ryu, K.B.; Park, K.R. OADE-Net: Original and attention-guided denseNet-based ensemble network for person re-identification using infrared light images. *Mathematics* **2022**, *10*, 3503. [CrossRef]
8. Zhang, X.; Wang, Y.; Peng, X.; Dong, M.; Ota, K.; Xu, L. Ohma: An edge-based lightweight occluded target re-identification framework for exploring abundant feature expression. *IEEE Trans. Consum. Electron.* **2024**, *70*, 7424–7435. [CrossRef]
9. Ma, B.; Wu, Z.; Ge, Y.; Chen, B.; Lin, J.; Zhang, H.; Xia, H. Continuous picking path planning based on lightweight marigold corollas recognition in the field. *Biomimetics* **2025**, *10*, 648. [CrossRef] [PubMed]
10. Itti, L.; Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2001**, *2*, 194–203. [CrossRef] [PubMed]
11. Zhu, H.; Budhwant, P.; Zheng, Z.; Nevatia, R. SEAS: Shape-aligned supervision for person re-identification. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 164–174. [CrossRef]
12. Chen, Y.; Zou, G.; Huang, Z.; Chen, G.; Hu, B. Person re-identification based on multi-feature fusion to enhance pedestrian features. *Displays* **2026**, *91*, 103187. [CrossRef]
13. Mamedov, T.; Konushin, A.; Konushin, V. Remix: Training generalized person re-identification on a mixture of data. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 28 February–4 March 2025; pp. 8186–8196. [CrossRef]
14. Ning, E.; Miao, J.; Xie, S.; Ma, H.; Ning, X. Occluded person re-identification in multi-scenarios: A synergistic interaction framework with perception-aware optimization. *Eng. Appl. Artif. Intell.* **2026**, *166*, 113674. [CrossRef]
15. Ding, Y.; Wang, Y.; Jian, Y.; Yu, C.; Tian, K.; Ye, Z.; Jiang, G. FA-ReID: Feature alignment with adversarial learning for clothing-changing person re-identification. *Neurocomputing* **2026**, *666*, 132237. [CrossRef]
16. Zhang, Y.; Kang, W.; Liu, Y.; Zhu, P. Multi-scale semantic and detail extraction network for lightweight person re-identification. *Comput. Vis. Image Underst.* **2023**, *236*, 103813. [CrossRef]
17. Wang, H.; Sun, Y.; Bi, X. Structural redundancy reduction based efficient training for lightweight person re-identification. *Inf. Sci.* **2023**, *637*, 118962. [CrossRef]
18. Jin, W.; Yanbin, D.; Haiming, C. Lightweight person re-identification for edge computing. *IEEE Access* **2024**, *12*, 75899–75906. [CrossRef]

19. Liu, Q.; Ma, M.; Shen, P.; Hu, B.; Yuan, X.; Guo, T.; Gao, M. A lightweight multi-scale feature enhancement network for person re-id. *Expert Syst.* **2025**, *42*, e70165. [CrossRef]
20. Zhang, Y.; Qin, Y. Lightweight person re-identification network based on random color dropout and self-attention. *Tsinghua Sci. Technol.* **2026**, *31*, 976–992. [CrossRef]
21. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3701–3711. [CrossRef]
22. Cui, Z.; Zhou, J.; Peng, Y. DMA: Dual modality-aware alignment for visible-infrared person re-identification. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 2696–2708. [CrossRef]
23. Qiu, L.; Chen, S.; Xue, J.H.; Wang, D.H.; Zhu, S.; Yan, Y. Hoh-net: High-order hierarchical middle-feature learning network for visible-infrared person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2026**, *36*, 2607–2622. [CrossRef]
24. Yang, X.; Liu, H.; Wang, N.; Gao, X. Bidirectional modality information interaction for visible–infrared person re-identification. *Pattern Recognit.* **2025**, *161*, 111301. [CrossRef]
25. Yu, X.; Dong, N.; Zhu, L.; Peng, H.; Tao, D. Clip-driven semantic discovery network for visible-infrared person re-identification. *IEEE Trans. Multimed.* **2025**, *27*, 4137–4150. [CrossRef]
26. Zhang, Y.; Wang, H.; Lu, Y.; Yan, Y.; Li, X. Frequency domain nuances mining for visible-infrared person re-identification. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 5411–5424. [CrossRef]
27. Yu, H.; Cheng, X.; Peng, W. Toplight: Lightweight neural networks with task-oriented pretraining for visible-infrared recognition. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 3541–3550. [CrossRef]
28. Josi, A.; Alehdaghi, M.; Cruz, R.M.O.; Granger, E. Multimodal data augmentation for visual-infrared person reid with corrupted data. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 3–7 January 2023; pp. 1–10. [CrossRef]
29. Deng, S.; Yuan, K.; Schaefer, G.; Zhou, S.; Vogiatzis, G.; Wang, Y.; Fang, H. A dual-aligned knowledge self-distillation framework for visible-infrared cross-modal person re-identification. *Knowl.-Based Syst.* **2025**, *330*, 114525. [CrossRef]
30. Zhang, J.; Yuan, Y.; Wang, Q. Night person re-Identification and a benchmark. *IEEE Access* **2019**, *7*, 95496–95504. [CrossRef]
31. Jeong, M.S.; Jeong, S.I.; Lee, D.C.; Jung, S.Y.; Park, K.R. Weak saliency ensemble network for person re-identification using infrared light images. *Eng. Appl. Artif. Intell.* **2025**, *139*, 109517. [CrossRef]
32. Kim, J.; Bhalgat, Y.; Lee, J.; Patel, C.; Kwak, N. QKD: Quantization-aware knowledge distillation. *arXiv* **2019**, arXiv:1911.12491. [CrossRef]
33. Zhao, K.; Zhao, M. Self-supervised quantization-aware knowledge distillation. In Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS), Valencia, Spain, 2–4 May 2024; pp. 4375–4383. Available online: https://proceedings.mlr.press/v238/zhao24d.html (accessed on 15 November 2026).
34. Kim, M.; Lee, S.; Hong, S.; Chang, D.-S.; Choi, J. Understanding and improving knowledge distillation for quantization aware training of large transformer encoders. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 6713–6725. [CrossRef]
35. Herzog, F.; Ji, X.; Teepe, T.; Hormann, S.; Gilg, J.; Rigoll, G. Lightweight multi-branch network for person re-identification. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 1129–1133. [CrossRef]
36. Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; Gu, J. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Trans. Multimed.* **2020**, *22*, 2597–2609. [CrossRef]
37. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826. [CrossRef]
38. Wang, X.; Han, X.; Huang, W.; Dong, D.; Scott, M.R. Multi-similarity loss with general pair weighting for deep metric Learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5017–5025. [CrossRef]
39. Nguyen, D.T.; Hong, H.G.; Kim, K.W.; Park, K.R. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* **2017**, *17*, 605. [CrossRef] [PubMed]
40. Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; Lai, J. RGB-infrared cross-modality person re-identification. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5390–5399. [CrossRef]
41. Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; Tang, J. Robust multi-modality person re-identification. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI), Virtual, 2–9 February 2021; pp. 3529–3537. [CrossRef]

42. Zhang, K.; Liang, J.; Van Gool, L.; Timofte, R. Designing a practical degradation model for deep blind image super-resolution. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 4771–4780. [CrossRef]
43. NVIDIA. GeForce RTX 20 Series Graphics Cards. Available online: https://www.nvidia.com/en-us/geforce/20-series/ (accessed on 2 May 2026).
44. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 501–518. [CrossRef]
45. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based object re-identification. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 14993–15002. [CrossRef]
46. Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; Li, Y. Maxvit: Multi-axis vision transformer. In Proceedings of the 2022 European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 459–479. [CrossRef]
47. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11999–12009. [CrossRef]
48. Wang, H.; Shen, J.; Liu, Y.; Gao, Y.; Gavves, E. Nformer: Robust person re-identification with neighbor transformer. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 7287–7297. [CrossRef]
49. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 16133–16142. [CrossRef]
50. Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; Fu, Y. Rewrite the stars. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 5694–5703. [CrossRef]
51. Qin, D.; Leichner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. Mobilenetv4: Universal models for the mobile ecosystem. In Proceedings of the 2024 European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 78–96. [CrossRef]
52. Munir, M.; Rahman, M.M.; Marculescu, R. Rapidnet: Multi-level dilated convolution based mobile backbone. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 28 February–4 March 2025; pp. 8302–8312. [CrossRef]
53. Hatamizadeh, A.; Kautz, J. Mambavision: A hybrid mamba-transformer vision backbone. In Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 25261–25270. [CrossRef]
54. Mishra, P.; Singh, U.; Pandey, C.M.; Mishra, P.; Pandey, G. Application of student's *t*-test, analysis of variance, and covariance. *Ann. Card. Anaesth.* **2019**, *22*, 407–411. [CrossRef] [PubMed]
55. Cohen, J. A power primer. In *Methodological Issues and Strategies in Clinical Research*, 4th ed.; American Psychological Association: Washington, DC, USA, 2016; pp. 279–284. [CrossRef]
56. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
57. Mahmood, T.; Wahid, A.; Hong, J.S.; Kim, S.G.; Park, K.R. A Novel Convolution Transformer-based Network for Histopathology-image Classification Using Adaptive Convolution and Dynamic Attention. *Eng. Appl. Artif. Intell.* **2024**, *135*, 108824. [CrossRef]
58. Haider, A.; Arsalan, M.; Park, C.; Sultan, H.; Park, K.R. Exploring Deep Feature-Blending Capabilities to Assist Glaucoma Screening. *Appl. Soft Comput.* **2023**, *133*, 109918. [CrossRef]
59. Kim, Y.H.; Nam, S.H.; Hong, S.B.; Park, K.R. GRA-GAN: Generative Adversarial Network for Image Style Transfer of Gender, Race, and Age. *Expert Syst. Appl.* **2022**, *198*, 116792. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.