Learning Resolution-Adaptive Representations for Cross-Resolution Person Re-Identification

Lin Yuanbo Wu, Lingqiao Liu, Yang Wang, Zheng Zhang, Farid Boussaid, Mohammed Bennamoun, Xianghua Xie

***Abstract*—Cross-resolution person re-identification (CRReID) is a challenging and practical problem that involves matching low-resolution (LR) query identity images against high-resolution (HR) gallery images. Query images often suffer from resolution degradation due to the different capturing conditions from real-world cameras. State-of-the-art solutions for CRReID either learn a resolution-invariant representation or adopt a super-resolution (SR) module to recover the missing information from the LR query. In this paper, we propose an alternative SR-free paradigm to directly compare HR and LR images via a dynamic metric that is adaptive to the resolution of a query image. We realize this idea by learning resolution-adaptive representations for cross-resolution comparison. We propose two resolution-adaptive mechanisms to achieve this. The first mechanism encodes the resolution specifics into different subvectors in the penultimate layer of the deep neural network, creating a varying-length representation. To better extract resolution-dependent information, we further propose to learn resolution-adaptive masks for intermediate residual feature blocks. A novel progressive learning strategy is proposed to train those masks properly. These two mechanisms are combined to boost the performance of CRReID. Experimental results show that the proposed method outperforms existing approaches and achieves state-of-the-art performance on multiple CRReID benchmarks.**

***Index Terms*—Cross resolution person re-identification, resolution-adaptive representations, resolution-adaptive masking.**

I. Introduction

PERSON re-identification (re-ID) is a critical task that involves matching the image of the same person across different images captured by various cameras. The task is gaining increasing attention due to its wide range of applications in person tracking [1], surveillance, and forensics [2]. Existing works in re-ID focus on developing feature representations or metrics that can handle image variations due to illumination changes or occlusions [3], [4]. However, all these methods assume that both the query and gallery images have similar high resolutions. In real-world scenarios, this assumption may not hold true, as image resolution may vary due to different distances between cameras and the subject of interest. For example, images captured by surveillance cameras (i.e., the query image) are generally in low resolution (LR), whereas the gallery images are typically in high resolution (HR). Directly matching an LR query against an HR gallery usually leads to inferior performance. This gives rise to the problem of cross-resolution person re-identification (CRReID).

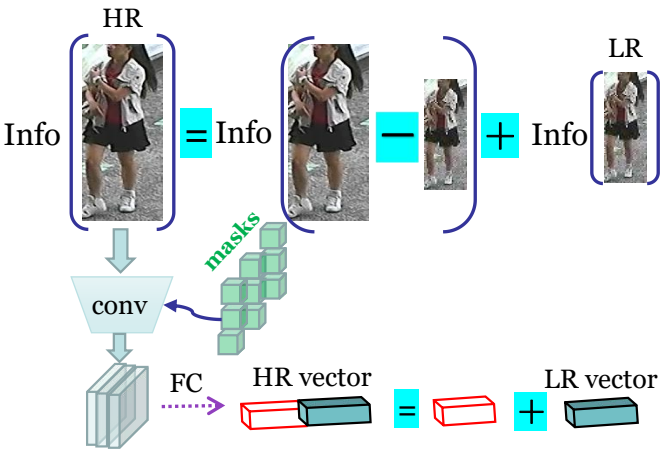


Fig. 1: The proposed method learns resolution-adaptive representations for CRReID. Through a resolution-adaptive masking and a varying-length representation, we can encode an HR image into the sub-vectors corresponding to the shared information with its LR counterpart, and more sub-vectors corresponding to the extra HR information.

To address the CRReID problem, state-of-the-art (SOTA) methods would employ either methods [5]–[9] with super-resolution (SR) modules or methods [5], [6], [10] that learn resolution-invariant features. The former first recovers the missing details of LR queries before performing the CRReID practice. The basic assumption is that by using the prior knowledge learned from the training data, the missing details of LR images can be recovered or at least be estimated in a way that will benefit the cross-resolution comparison. However, such a pipeline heavily depends on the recovery outcome, and yet there is no guarantee that useful details can be recovered. Moreover, if the input resolution is not seen by the SR model, one cannot properly recover the HR details for unseen resolution. The latter line of research on resolution variance attempts to learn feature representation that are invariant to resolutions so as to facilitate the cross-resolution comparison. However, such a scheme might have the risk of losing resolution specifics due to the invariant

L. Wu is with Hefei University of Technology, China and also with Swansea University, SA1 8EN, United Kingdom. E-mail: jolin.lwu@gmailcom.
L. Liu is with University of Adelaide, Adelaide 5005, Australia. E-mail: lingqiao.liu@adelaide.edu.au.
F. Boussaid and M. Bennamoun are with The University of Western Australia, Perth 6009, Australia. E-mail: {Farid.Boussaid;Mohammed.Bennamoun}@uwa.edu.au.
Y. Wang is with Hefei Univerity of Technology, Hefei 230009, China. E-mail: yangwang@hfut.edu.cn.
Z. Zhang is with School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China. E-mail: darrenzz219@gmail.com.
X. Xie is with Department of Computer Science, Swansea University, SA1 8EN, United Kingdom. E-mail: x.xie@swansea.ac.uk.

enforcement.

This paper presents a novel approach to compare HR and LR images without relying on super-resolution modules or invariant features. The proposed method aims to build a metric that is adaptive to the resolution of the query image, allowing it to select the most appropriate distance metric to compare with the HR gallery images. This is achieved through learning resolution-adaptive representations, as illustrated in Fig. 1. Specifically, the approach involves two mechanisms, both of which are discussed in Section III-B and III-C. The first mechanism is a varying-length image representation, where the representation length is determined by the resolution of the image. This encourages the explicit encoding of shared features across resolution and HR-specific features. The representation of an HR image is composed of sub-vectors corresponding to its LR counterpart and extra dimensions corresponding to the higher resolution part. This design captures the essential relationship between HR and LR images and enables images with different resolutions to be compared via their shared sub-vectors. The second mechanism further enhances the resolution-adaptive capability by learning resolution-specific masks that are applied to the intermediate activation of a neural network. Since there is a resolution-dependent correlation between the feature blocks, co-adaptation may occur if multiple masks are trained end-to-end [11]. Therefore, we develop a progressive training strategy that is demonstrated to be more effective than the standard end-to-end training for these resolution-adaptive masks. Through extensive experiments, we show that the two proposed mechanisms are complementary to be combined to achieve superior performance to state-of-the-art approaches that solely rely on super-resolution techniques (see Section V).

Our contributions are summarized below.

- We propose a varying-length representation that can adaptively encode the visual patterns of images from different resolutions. This enables convenient comparisons between images at different resolutions.
- We design a resolution-adaptive mechanism by introducing resolution-adaptive masks for intermediate residual feature blocks.
- We propose a novel progressive training strategy for training a group of resolution-adaptive masks. This can effectively combat the co-adaptation circumstance.

II. Related Work

A. Standard Person Re-ID

Recent person re-ID methods provide person representations that are robust to variations caused by various factors such as human pose, occlusion, and background clutter [12]–[15]. For instance, part-based methods [16]–[24] describe a person image as a combination of body parts either explicitly or implicitly. A number of explicit part-based methods use off-the-shelf pose estimators to extract body parts (e.g., head, torso, legs) with their corresponding features. Instead of explicitly estimating the human pose, implicit part-based methods [16], [22], [25] rather divide each person image into different horizontal parts with multiple scales [26]. As such, they can exploit the various partial information of the image, and provide a feature representation that is robust to occlusion. Several other approaches [21], [27]–[29] leverage attention mechanisms to highlight the discriminative parts and remove the background clutter. Other research directions focus on using domain adaptation for person re-ID [30]–[33]. For instance, Zhong *et al.* [31] proposed to generalize the re-ID model by considering the intra-domain variations of the target domain. Bai *et al.* [33] improved unsupervised domain adaptation (UDA) for re-ID by identifying the domain-specific and domain-fusion views. However, all aforementioned approaches assume that both query and gallery images have similar (high) resolutions, making them not suitable to real-world scenarios.

B. Cross-Resolution Person Re-ID (CRReID)

To address the practical challenge of CRReID, two main categories of methods have been developed: 1) metric learning or dictionary learning based approaches [34]–[36]; and 2) super-resolution (SR) based approaches [5]–[9], [37], [38]. For instance, to overcome the resolution mismatch, Jing *et al.* [34] developed a semi-coupled low-rank dictionary learning method to associate the mapping between the HR and LR images. Li *et al.* [35] introduced a method to jointly perform the cross-scale image alignment and multi-scale distance metric learning. However, all above methods are inherently limited in their matching ability due to the missing details in LR images.

Super-resolution based approaches cope with cross-resolution re-ID via a recovery and re-ID process [39]. An early work presented by Jiao *et al.* [5] uses a set of SR sub-networks to improve the compatibility between SR and re-ID. CSR-Net [40] explores cascading multiple SR-GANs [41] to progressively recover the details of LR images for resolution alignment. However, these models adopt a separate SR component in the recovery and re-ID pipeline. Instead of applying separate SR models, a novel architecture based on adversarial learning, called RAIN [10], was proposed to align and extract the resolution-invariant features in an end-to-end fashion. Inspired by RAIN [10], CAD-Net [6] further improves the performance by aligning the distributions between HR and LR images. More specifically, CAD-Net [6] jointly considers the resolution-invariant representations and the fine-grained detail recovery in LR input images. However, an outstanding issue remains is that the resolution of the query is unknown to us. To tackle this issue, Han *et al.* [7] proposed a framework to adaptively predict the optimum scale factor for the LR images so as to benefit the recovery for the CRReID.

Another recent work presented by Cheng *et al.* [8] explores the influence of different resolutions on feature extraction. They show that LR images can provide complementary information to the HR images. Considering the complementary information from the LR images, Cheng *et al.* [8] developed a joint multi-resolution framework based on a reconstruction network and a multi-branch feature fusion network. Following that, multi-resolution features are fully utilized in feature extraction by using channel-attention or residual Transformer blocks [9], [42]. In contrast, this paper proposes a method that has its own differences: first, our method does not need

excessive feature extraction for cross-resolution matching; and second we directly learn resolution-adaptive representations which are amenable for cross-resolution comparison.

III. Proposed Method

In this section, we first provide a problem statement and an algorithmic overview of our approach, and then elaborate on components of the proposed method. Central to our framework are two mechanisms: **(1)** a varying-length representation learning to encode the resolution-specific information into different sub-vectors; and **(2)** a set of learnable resolution-adaptive masks that are applied to intermediate feature blocks at different residual convolutions.

A. Problem Statement and Framework Overview

We aim to learn a model M that can match a low-resolution query image against the high-resolution gallery images. We assume that the resolutions of both the query and gallery images are provided. In practice, the resolution could be estimated from the size (number of pixels) of images or pedestrian bounding boxes since the height of people is relatively fixed. Without loss of generality, we assume that the resolution could be quantized into a set of discrete levels. We denote the resolution with k, e.g., $k \in \{1, 1/2, 1/3, 1/4\}$, which is the proportion of the height/width dimension as opposed to the highest resolution considered ($k = 1$ refers to the highest resolution). For example, if the highest resolution corresponds to 256×128 per person image, i.e., its resolution ratio $k = 1$, for a LR image of size 64×32, its resolution ratio becomes $k = 1/4$. In our algorithm, we resize all the images, whether LR or HR, to equal to the size of the highest resolution images via bilinear up-sampling. Then, each full-resolution image is down-sampled at different specified ratios to form its LR alternations.

Specifically, following the setting of CRReID [5], we down-sample the HR training images to form various LR images to simulate the LR query images. The aim of our training algorithm is to learn a resolution-adaptive metric, that is:

$$dist(x_p, x_g) = M(x_p, x_g, k), \tag{1}$$

where $dist(x_p, x_g)$ returns the distance between a query image x_p and a gallery x_g. We implement this similarity measure via a learnable model M. An important characteristic of this model is that the resolution ratio of the query image k is the input of M, and thus the metric is *resolution adaptive*. More specifically, we implement M by *learning resolution-adaptive representations* and we propose two resolution-adaptive mechanisms to realize that. The first is a varying-length representation that uses varying dimensions to encode a query image with different resolutions: the higher resolution, the longer dimensionality of the representation. We hypothesize that the representation of a higher resolution image should contain the common fragmented dimension shared with a LR image and additional dimensions which depict its own extra information. To further extract resolution-specific information, we propose the second mechanism, i.e., injecting resolution-specific masks into the intermediate residual feature blocks. This strongly enhances the resolution adaptive capability of the network via resolution-dependent mask generation. We depict the two mechanisms in Fig. 2, and details are presented in Section III-B and III-C, respectively.

Discussions Comparing with the resolution-adaptive metric, the resolution-invariant metric or representation seems to be a viable solution. However, since the resolution of the query image is not fixed, learning resolution-invariant features will identify discriminative information that are *shared across all resolutions*. Consequently, the information specific to resolutions higher than the lowest one will not be preserved. This inevitably prevent the network from using more information for matching a moderate LR query to HR gallery images.

B. Mechanism 1: Learning Varying-Length Resolution-Adaptive Representations

The varying-length resolution-adaptive representation is motivated by the relationship between HR images and LR images: *a HR image should contain all the information conveyed in the LR image, but also extra information from the higher resolution*. Therefore, when comparing a LR image with a HR image, the comparison should adhere to the information shared between them. Note that we do not assume that the unobserved high-resolution information can be recovered from a LR image by leveraging the object prior as the most super-resolution-based CRReID methods. Applying the above idea into representation learning, we propose to encode the information shared across resolutions and the information specific to HR into different dimensions of the feature representation. For example, for a HR image and a LR image, their shared part will be encoded into a sub-vector of the feature representation and the HR-specific part should be encoded into another sub-vector. When one compares a HR image and a LR image, the comparison should only be based on the shared part. In other words, for a fixed-size representation, the LR image will be encoded into the bits with lower dimensions (shorter length). In CRReID, a query image could have different resolutions, thus the above strategy will result in different representation lengths, i.e., the higher resolution of the query is, the more information that can be shared with the HR gallery images, and thus the longer dimension of the representation is.

In our implementation, we define m sub-vectors $\{\mathbf{v}_k\}, k = 1, \cdots, m$, with m corresponding to m different levels of resolution. For images with the highest resolution, all m sub-vectors will be concatenated as the final image representation. For the lowest resolution, only the first sub-vector will be activated. Formally, the representation of a query image that corresponds to the k-th level resolution, (larger k, higher resolution), is $\mathbf{z}_p = cat(\mathbf{v}_1^p, \cdots, \mathbf{v}_k^p, \mathbf{v}_{k+1:m}^p)$, where where $cat(\cdot)$ denotes concatenation. Since $\mathbf{z}_p$ is at resolution k, the sub-vectors $\mathbf{v}_{k+1:m}^p$ are zeros. Thus, we have the truncated $\hat{\mathbf{z}}_p = cat(\mathbf{v}_{1:k}^p)$ in short. For HR gallery images, their representations are the concatenations of all m sub-vectors, that is, $\mathbf{z}_g = cat(\mathbf{v}_1^g, \cdots, \mathbf{v}_k^g, \ldots, \mathbf{v}_m^g)$. When a level-$k$-resolution query image x_p is compared against a HR gallery image x_g, the distance is calculated via

$$dis(x_p, x_g) = \|\mathbf{z}_p - \mathbf{z}_g\|_2^2 = \|\hat{\mathbf{z}}_p - \hat{\mathbf{z}}_g\|_2^2, \tag{2}$$

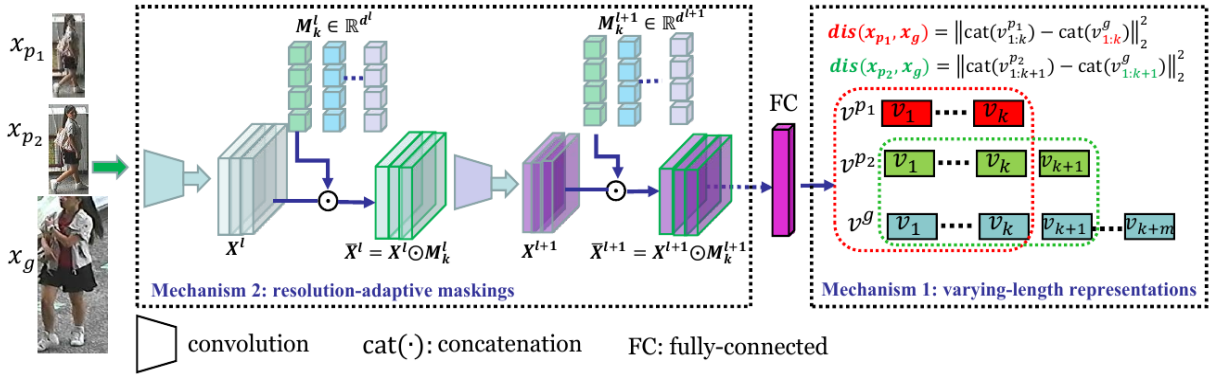


Fig. 2: The scheme of learning resolution-adaptive representations for person images with different resolutions. Given the query images $x_{p_i}(i = 1, 2)$, and a gallery image x_g at different resolution levels, we propose two mechanisms to learn resolution-adaptive representations that are convenient for cross-resolution comparison. **Mechanism 1**: the varying-length representation learning to produce a feature vector with varied dimensions, i.e., the gallery vector v^g contains all sub-vectors that the query vector v^{p_i} has, plus extra dimensions in the HR image. **Mechanism 2**: resolution-adaptive maskings ($\mathbf{M}^l$) are applied to the l-th feature block $\mathbf{X}^l$, yielding $\bar{\mathbf{X}}^l$. The final resolution-adaptive metric is returned by calculating the distance: $dis(x_{p_i}, x_g)$. Figure best viewed in color.

where $\hat{\mathbf{z}}_g = cat(\mathbf{v}_{1:k}^g)$. In other words, the comparison is conducted by only comparing the top-k sub-vectors of $\mathbf{z}_g$ when we know the query image resolution is at level-k.

C. Mechanism 2: Resolution-Adaptive Masking

The above varying-length representation only adaptively constructs the resolution-specific representation in the penultimate layer of the neural network. To extract more resolution-dependent features, we propose a mechanism to inject the resolution characteristics into the earlier layers of a neural network. More specifically, we build our network based on a residual network [43] with learnable masks: one for a resolution level to the activations after each residual block. Each mask is a vector, with each dimension being a real value between 0 and 1. The mask acts as a dimension-wise scaling factor to the feature maps. Formally, let $\mathbf{X}^l \in \mathbb{R}^{d^l \times H^l \times W^l}$ denote the feature maps after the l-th residual block. A set of masks $\{\mathbf{M}_k^l \in \mathbb{R}^{d^l}\}, k = 1, \cdots, m$, are applied to $\mathbf{X}^l$ by $\bar{\mathbf{X}}^l = \mathbf{X}^l \odot \mathbf{M}_k^l$ [1], where $\odot$ denotes the element-wise multiplication and k is the resolution-level of the input image. For input images with varied resolutions, different masks will be chosen to determine the final representation. The values of $\mathbf{M}_k^l$ are parameters to be learned. In practice, we reformulate those masks as an channel-wise scaling layer and learn the layer parameters with the network:

$$\bar{\mathbf{X}}^l = \mathbf{X}^l \odot \big(\sum_k s_k^l \text{Sigmoid}(\mathbf{M}_k^l)\big), \tag{3}$$

where $s_k^l = 1$ if the input image is at resolution level k, otherwise $s_k^l = 0$. s^l could be considered as an input to the network. Sigmoid is the Sigmoid function converts the real-valued layer parameters $\mathbf{M}_k^l$ into the range between 0 and 1.

[1]Please note that we DO NOT have any constraints on the structure of those masks, e.g., requiring each dimension of an individual mask corresponding to certain resolutions. For a given layer, we simply allocate one mask for each level of resolution.

Each column $\mathbf{M}_k^l[i]$ in the matrix $\mathbf{M}_k^l$ represents a mask at each resolution level k. These masks are not binary, but instead use real-valued scaling coefficients that are applied to the feature tensor $\mathbf{X}^l$ at the l-the residual block. Being masked with respect to a specific resolution, the network is guided to producing more more resolution-adaptive features so as to enrich the representation capacity. This operation incurs no additional training cost. It is important to note that the soft masks (i.e., $\mathbf{M}_k^l$ and $\mathbf{M}_k^{l+1}$) at different blocks are not weight-shared. Each block-wise mask accounts for features with increasing complexity and is trainable, making it possible to jointly learn them end-to-end. We recall that developing mask generators is equivalent to aligning person images with occlusion, wherein visible patterns from non-occluded images can be selected by corresponding masks to align and compare with occluded regions [3], [4]. It is worth noting that our proposed resolution-dependent masks are applied in a channel-wise manner to reflect the resolution levels in feature dimensions, making them suitable for CRReID.

D. Varying-length Sub-vectors with Resolution Variations

To enable direct comparison between images at different resolutions, we propose a varying-length feature that reflects the query resolution. Given a LR query image, it is encoded into a feature vector with resolution-dependent dimension. Since the LR image shares content with the original HR image but also contains its own characteristics, the feature vector of each image should be a combination of commonality and resolution-induced characteristics. However, a deep feature representation outputted from neural networks has a fixed-size dimension, making it challenging to define varied feature dimension corresponding to different resolution levels. To overcome this challenge, we propose to predict a set of sub-vectors, where the number of sub-vectors corresponds to the resolution. This induces the idea of varying-length feature.

Algorithm 1 Resolution Adaptive Representation Learning for CRRe-ID.

Inputs: Training tuple in the form of query image x_p, gallery image x_g, and the resolution level of the query image k: $\{x_p, x_g, k\}$.
Output: The trainable model M that learns varying-length features for both x_p and x_g.

1: Define C identity classifiers, $\mathbf{W} \in \mathbb{R}^{d\times C}$, one for each identity.
2: **for** $l = 1, 2, \cdots, L$ **do**
3: Randomly initialize the layer-wise masks $\mathbf{M}^l = \{\mathbf{M}^l_k\}$ at the l-th layer, and fix the parameters of $\mathbf{M}^{l'}$,$\forall l' < l$. (The lower layer index, the closer to the output layer).
4: **for** $t = 1, 2, \cdots T$ **do**
5: 1. Randomly sample a mini-batch of training triplets. Each sample is with the form $\{x_p, x_g, k\}$.
6: 2. For each triplet, with regards to the query resolution k and the current layer l, determine s^l_k for the scaling layer in Eq. (3). Perform forward calculation to obtain varying-length representation $\mathbf{z}_k = cat(\mathbf{v}_{1:k})$.
7: 3. Padding zero to $\mathbf{z}_k$, making its dimension equal to d. Apply both class identity loss and verification loss via Eq. (4). Then perform back-propagation.
8: **end for**
9: **end for**

Specifically, we train a classifier consisting of a set of sub-classifiers, such that an image at a resolution looks up the sub-classifiers to adaptively characterise its own features. Consider an image x with its resolution indication k, its deep feature vector $\mathbf{v} = [\mathbf{v}_1, \ldots, \mathbf{v}_k, \ldots, \mathbf{v}_m] \in \mathbb{R}^d$ is outputted from the last fully-connected layer of the network, with d denoting the dimensionality and $\mathbf{v}_k \in \mathbb{R}^{d_k}$ is a partition of $\mathbf{v}$ with the dimensionality d_k. We further define $\mathbf{w}^i$ as a classifier for one identity i, then for all identities ($i = 1, \ldots, C$), we have a set of classifiers $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^k, \ldots, \mathbf{w}^C] \in \mathbb{R}^{d\times C}$, where $\mathbf{w}^k \in \mathbb{R}^{d_k\times C}$ is a prototype-based sub-classifier, and C is the number of total identities. To perform the prediction on the varying-length feature vector $\mathbf{v}$, we calculate the identity prediction logits across all identities via $\mathbf{e}^k = (\mathbf{w}^k)^T\mathbf{v}_k \in \mathbb{R}^C$. Since the identity prediction classifies each image by evaluating the classifier $\mathbf{w}^k$ into the embedding space, the classifier can be interpreted as the prototype closest to the image in the feature space. An image is assigned to the identity label of its nearest prototype. Thus, the prototype-based classifier is effective for classifying images based on the closest training prototypes $\mathbf{w}^k$ in the feature space. Finally, the learning objective yielded by the varying-length prediction incrementally updates prototypes to better discriminate the training images with identity labels. The computational algorithm is illustrated in Algorithm 1.

E. Resolution-Adaptive Representation Training

Most of the state-of-the-art (SOTA) person re-ID methods train the models with an identity classification loss $\mathcal{L}_{\text{cls}}$ (namely ID loss) and a verification loss $\mathcal{L}_{\text{verif}}$. We follow this convention to adopt both losses to train the proposed model. In traditional re-ID model training, the standard ID loss is applied to the fixed-length representation. However, our training set comprises HR images and multiple LR counterparts, which are created by down-sampling the HR images with varied resolution levels. This leads to multiple representations when applying the resolution-adaptive mechanisms. To overcome this issue, we propose to apply zero-padding, i.e., concatenating "0" to the representation whose dimension is less than the maximal dimension, to convert a varying-length representation to a fixed-length representation. Then a normal identity classification loss can be applied to learn the model.

The verification loss $\mathcal{L}_{\text{verif}}$ is applied to a binary classifier that predicts whether two samples belong to the same class. In our implementation, we start by padding zeros to the varying-length representations of two images, and send their feature vector difference to a multi-layer perceptron (MLP) to make a binary prediction about whether those two samples are from the same class. The loss function is

$$\begin{aligned}\mathcal{L}_{\text{verif}} = -\sum_{n}^{N} y_n \log(p(y_n = 1 \mid \mathbf{v}_{ij})) \\ + (1 - y_n)\log(1 - p(y_n = 1 \mid \mathbf{v}_{ij})),\end{aligned} \tag{4}$$

where $\mathbf{v}_{ij} = \mathbf{v}_i - \mathbf{v}_j$ denotes the feature difference and $p(y_n = 1 \mid \mathbf{v}_{ij})$ is implemented with a MLP, e.g., $p(y_n = 1 \mid \mathbf{v}_{ij}) = Sigmoid(f(\mathbf{v}_{ij}))$, where f is a MLP mapping $\mathbf{v}_{ij}$ to a scalar. We define $y_n = 1$ if two images are from the same class, otherwise $y_n = 0$. Our final loss is the weighted summation of both loss terms, that is, $\mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{verif}}$, where λ is the balance parameter.

1) Analysis of the Identity Classification Loss: The verification loss takes inputs from two images from different or same resolutions. It can learn a resolution-adaptive metric naturally. However, the identity classification loss only uses one image as input. One may wonder if it can also be beneficial for resolution-adaptive metric learning? To answer this question, we conduct an analysis in the following part. In the following part, we present an analysis on the identity classification loss to show why it can benefit the resolution-adaptive metric learning. Note that the inner product between an identity classifier and a zero-padded representation will only be determined by the inner product of the sub-vectors corresponding to the non-zero parts of the representation. Formally, we could consider that a classifier is constructed with m parts too, each part corresponding to one sub-vector in the varying-length representation. In other words, $\mathbf{w} = cat(\mathbf{w}^1, \ldots, \mathbf{w}^m) \in \mathbb{R}^d$, where $\mathbf{w}^k \in \mathbb{R}^{d_k}$ and $cat(\cdot)$ represents concatenation. As a result, we have $\mathbf{w}^\top\mathbf{z}_k = cat(\mathbf{w}^1, \ldots, \mathbf{w}^k)^\top\mathbf{z}_k$, where $\mathbf{z}_k$ is an image representation with resolution level k. The ID loss will encourage samples from the same identity class to move closer to the corresponding classifier $\mathbf{w}$ and thus indirectly pulling those features close to each other. Similarly, we could expect our ID loss will pull $\mathbf{z}_k$ and the first k-th sub-vectors of $\mathbf{z}_{k'}$ $k' > k$ close to each other, which ensures that images of the same identity but different resolutions become closer under the proposed distance metric Eq. (2). This can be explained by the following example.

In the ID loss, each identity classifier can be interpreted as the prototype for each identity. When we take the inner product between a prototype and a feature vector at the k-th level resolution, only the first k sub-vectors of the prototype are used for comparison. By concatenating the first k sub-vectors, we create a prototype for the images at resolution level k. The ID loss encourages image representations from the same

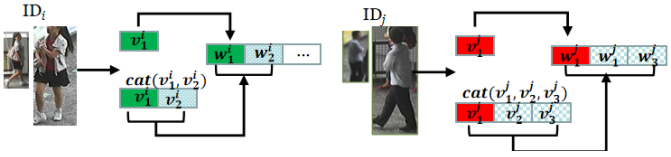


Fig. 3: By zero-padding, images with the same identity but different resolutions will lookup the same classifier. This will encourage the shared parts (denoted as the red and green blocks in different identities) of an HR-image representation and an LR-image representation close to each other if they belong to the same identity.

identity class to move closer to the identity prototype, making them more similar to each other. Similarly, training with varying-length representations encourages a LR representation to match the low resolution part of a higher resolution representation, which is beneficial for cross-resolution comparison. This can be seen in Example 1. Therefore, by training varying-length classifiers with the ID loss, we achieve an equivalent effect for learning representations at different resolutions. Additionally, learning both varying-length representations and classifiers encourages a LR representation to match the low part of a higher resolution representation, further benefiting cross-resolution comparison.

Example 1: Suppose an HR image and an LR image belonging to the same identity. Based on the varying-length representation, the HR image produces a representation $cat(\mathbf{v}_L^1, \mathbf{v}_H^1)$ while the LR image produces a representation $\mathbf{v}_L^2$. Assume the corresponding non-zero part of the classifier of this identity is $cat(\mathbf{w}_L, \mathbf{w}_H)$, then the ID loss can make $\mathbf{v}_L^2$ align with $\mathbf{w}_L$ and $cat(\mathbf{v}_L^1, \mathbf{v}_H^1)$ align with $cat(\mathbf{w}_L, \mathbf{w}_H)$. The latter usually implies that $\mathbf{v}_L^1$ should align with $\mathbf{w}_L$. Thus, this alignment indirectly encourages $\mathbf{v}_L^1$ to be aligned with $\mathbf{v}_L^2$ through their shared aligning target $\mathbf{w}_L$. This idea is illustrated in Fig.3.

2) Progressive Training for Resolution-Adaptive Masks: We aim to jointly train the two mechanisms for optimizing the performance at its best. While the most straightforward way is to train them end-to-end via stochastic gradient descent (SGD), we empirically find that this convention leads to compromised performance. This is partially ascribe to the difficulty of optimization over non-shareable masks. Moreover, those channel-wise masks at different layers are highly correlated, and training those masks becomes nontrivial due to the co-adaptation problem [11]. To combat this issue, we propose an effective progressive training scheme. Alternatively, we propose to inject the masks at different layers sequentially and train them progressively to avoid the co-adaptation of multiple masks [11]. In our implementation, we first fabricate the masks into the residual blocks closest to the classifier layer and then gradually multiply more masks into the residual blocks downwards to lower. Once new masks are weaved to the residual block, the masks that have been trained with previous higher-levels will be fixed and not updated anymore. Please refer to the experimental section for more discussion on the effect of this progressive training strategy. The whole training process is shown in Algorithm 1.

IV. Implementation Details

a) Backbone on ResNet-50: The model was implemented using PyTorch. We built the network based on the ResNet-50 architecture with four residual blocks. For all resolutions, we resized each image to $256 \times 128 \times 3$ and padded with 10 pixels with zero values. Then, we randomly cropped the image into a 256×128 rectangular images. Each image was flipped horizontally with 0.5 probability. The training batch size was 32. The learning rate was set to be 0.00035 and is decayed to 3.5×10^{-5} and 3.5×10^{-6} after 40 epochs and 70 epochs, respectively. This warm-up learning rate is shown to be effective to boostrap the network as suggested by [44]. We trained the network using ADAM optimizer with 120 epochs in total. The masking mechanism was applied to each residual block, which is a composition of two layers of 3×3 conv/batch norm/relu. The last stride of the residual block was set to 1 to achieve a feature map with a higher spatial size (16×8). For each block, we initialized a resolution mask matrix using Gaussian randomness, followed by a sigmoid function. The sigmoid function restricts the real values in the range of (0,1). The parameter λ was empirically set to be 0.5. Empirically, we found that the sigmoid function should be applied on the resolution masks after each updating of the masking weights. This can effectively regularize the soft mask learning to control the magnitude of masks. This was shown to improve the learning capacity [45]. The whole network was trained in end-to-end with Stochastic Gradient Descent (SGD). We also performed a progressive training, in which the new masks are sequentially injected into the residual blocks from higher layers to lowers.

b) Backbone on Transformer: Our network is also instantiated with Transformer architecture [46], using both Swin-B and Swin-L Transformer blocks with 4 stages, named Ours-Swin-B and Ours-Swin-L respectively, in experiments. Note that Swin-B has the model size and computation complexity similar to Vision Transformer (ViT-base) [47]. These stages jointly produce a hierarchical representation with the same feature map resolutions as ResNet-50. To produce masking generation, we apply channel-wise maskings for specific resolutions in the transformer blocks.

V. Experiments

In this section, we evaluate the proposed method on several benchmark datasets and compare against SOTA methods. We report both quantitative and qualitative results as well as ablation studies to thoroughly analyze our method.

A. Datasets

Following existing works [5]–[7], we adopt the multiple low-resolution (MLR) person re-ID evaluation setting on four datasets. The details of each dataset are described as follows.

- **CAVIR** [48] dataset is a real-world dataset composed of 1,220 images of 72 identities and two camera views. Following [5]–[7], we discard 22 identities that only appear in the closer camera. The remaining images of 50 identities are randomly and evenly divided into two halves for training and test.

- **MLR-CUHK03** [49] dataset contains over 14,000 images of 1,467 identities captured by 5 pairs of cameras. Following [6], [7], we adopt the 1,367/100 identities as the training/test split.
- **MLR-Market-1501** [50] dataset consists of 32,668 person images of 1,501 identities observed under six different camera views. The dataset is split into 12,936 training images of 751 identities and 19,732 testing images of the remaining 750 identities.
- **MLR-DukeMTMC-reid** [51]: This dataset was collected with eight different cameras and was originally proposed for video-based person tracking and re-identification. It has a total of 1,404 identities and includes 16,522 training images of 702 identities, 2,228 query images of the other 702 identities and 17,661 gallery images.

Note that the CAVIR is the only real-world dataset, while MLR-CUHK03, MLR-Market-1501 and MLR-DukeMTMC-reid are *three synthetic* benchmarks. To construct the synthetic MLR datasets, that are, MLR-CUHK03, MLR-DukeMTMC-reid, MLR-Market1501, we adopt the setting in [5]–[7] and down-sample HR images taken by one camera by randomly choosing a down-sampling rate $r \in \{2, 3, 4\}$, while the other images remain unchanged.

B. Training and Evaluation Protocols

To train a CRReID model, we constructed the training set using the original HR images as well as the down-sampled images with the down-sampling rate $r \in \{2, 3, 4\}$. For evaluation, we followed the MLR evaluation protocol suggested in [5]–[7]. Specifically, for each HR image from the test set, we randomly chose a down-sampling rate $r \in \{2, 3, 4\}$ and used the down-sampled images to construct a query set. The gallery images are all HR images with one randomly selected HR image per person. The random data splits were repeated 10 times and the average value was computed for every 10 trials. For the re-ID evaluation, we used the average Cumulative Match Characteristic (CMC) and reported the results at rank-1, 5 and 10. In Section V-F, we also investigated the generalization of the proposed method to unseen resolutions in the inference.

C. Comparison Methods

We compared the proposed method against three families of SOTA approaches. The first family comprises a number of CRReID methods based on resolution-invariant representation including SING [5], RAIN [10], and CAD-Net [6]. The second family is based on the super-resolution module, which is the mainstream approach in CRReID. Methods like CRGAN [38], PRI [7] and DGRL [42] fall into this category. The third family aims for enhancing the discriminant ability of the deeply embedded features for different resolution images. DGRL [42] and PS-HRNet [9] belong to this category. The fourth family comprises standard re-ID models including CamStyle [53], FD-GAN [54], PCB [55], PyrNet [56] and DDGAN [37]. For those methods, we directly quote the results from PRI [7], where the authors strictly reproduce the results by using the same training set as our method. For standard re-ID methods, the training set contains the HR images only.

D. Experimental Results

1) Main Quantitative Results: The comparison results of our method against SOTA re-ID methods on four benchmark datasets are reported in Table I. For a fair comparison, we do not combine our method with any pre/post processing, e.g., re-ranking [57] or part-pooling [16], even though these operations can bootstrap the re-ID performance further. From Table I, we can make the following observations:**(1)** Comparing with resolution invariant representation learning methods, i.e., SING [5], RAIN [10] and CAD-Net [6], our method (Ours-ResNet-50) provides notable improvement across all evaluation metrics. For instance, our method (Ours-ResNet-50) outperforms CAD-Net [6] (a leading recovery and re-ID method) by 20.8%, 7.1%, 6.4% and 6.3% at rank-1 on CAVIR, MLR-CUHK03, MLR-Market-1501 and MLR-DukeMTMC-reid, respectively. This clearly supports our claim on the advantages of resolution-adaptive representations. **(2)** We observe that our method (Ours-ResNet-50) can also achieve superior performance over super-resolution based approaches, as shown in the comparison against CRGAN [38] and PRI [7], which are also SOTA methods in CRReID. Table I shows that our method (Ours-ResNet-50) outperforms those competing methods by a notable margin. This demonstrates that adaptive representation learning is a promising paradigm to solve the CRReID problem. The most competing method is PS-HRNet [9], which challenges traditional super-resolution restoration on low-resolution images. Instead, PS-HRNet [9] flags the pathway of preserving multi-resolution features and explicitly minimizes the impact between the feature distribution between LR and HR images. Our method coincides PS-HRNet [9] in the senses of enhancing the semantic information and reducing the impact of resolution difference in cross-resolution scenario. In contrast to PS-HRNet [9], we learn discriminant features for images at different resolution levels via the proposed varying-length prediction based on prototype classifiers (as described in Section III-D). And the proposed masking mechanism can effectively alleviate the feature difference by learning resolution adaptive representations. Table I shows that our method (Ours-Swin-L) outperforms PS-HRNet [9] by 17.6%, 5.6%, 2.6% and 2.9% at rank-1 across four datasets. One possible reason is the end-to-end training to learn discriminant and resolution-adaptive features, whereas PS-HRNet [9] adopts a multi-stage training on semantic extraction and cross-resolution difference mitigation. The most possible reason for the exceptional performance of Ours-Swin-L is that Swin-L transformer implements a hierarchical representation that can merge the feature resolutions across layers. This is especially advantageous to our framework in the sense of learning to implant the various query resolutions into feature hierarchy and then integrate these features. While Ours-Swin-L performs exceptionally well across all datasets, it is worth noting that the computational complexity of Swin-L is considerably high. To address this concern, we conducted experiments using Ours-Swin-B, which has a model size that is half that of Swin-L. The results presented in Table I demonstrate that Ours-Swin-B outperforms all other competitors and is surpassed only by Ours-Swin-L. These findings suggest that Ours-Swin-

TABLE I: Comparison with the state-of-the-art models on four datasets (%). Note that "-" means that the results have not been reported. Best results are in boldface. Our proposed method consistently outperforms all existing methods.

Method	CAVIAR			MLR-CUHK03			MLR-Market-1501			MLR-DukeMTMC-reid		
	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
JUDEA [35]	22.0	60.1	80.8	26.2	58.0	73.4	-	-	-	-	-	-
SLD^2L [34]	18.4	44.8	61.2	-	-	-	-	-	-	-	-	-
SDF [36]	14.3	37.5	62.5	22.2	48.0	64.0	-	-	-	-	-	-
SING [5]	33.5	72.7	89.0	67.7	90.7	94.7	74.4	87.8	91.6	65.2	80.1	84.8
CSR-GAN [40]	34.7	72.5	87.4	71.3	92.1	97.4	76.4	88.5	91.9	67.6	81.4	85.1
RAIN [10]	42.0	77.3	89.6	78.9	97.3	98.7	-	-	-	-	-	-
CAD-Net [6]	42.8	76.2	91.5	82.1	97.4	98.8	83.7	92.7	95.8	75.6	86.7	89.6
PRI [7]	43.2	78.5	91.9	85.2	97.5	98.8	84.9	93.5	96.1	78.3	87.5	91.4
INTACT [52]	-	-	-	86.4	97.4	98.1	88.1	95.0	96.4	81.2	90.1	92.8
CRGAN [38]	43.1	76.5	92.3	83.4	98.1	99.1	88.1	95.0	96.4	-	-	-
HRNet-ReID [9]	48.2	84.0	96.2	88.9	96.4	98.7	87.6	95.1	97.0	-	-	-
PS-HRNet [9]	48.2	84.5	96.3	92.6	98.3	99.4	91.5	96.7	97.9	82.3	90.5	92.8
CamStyle [53]	32.1	72.3	85.9	69.1	89.6	93.9	74.5	88.6	93.0	64.0	78.1	84.4
FD-GAN [54]	33.5	71.4	86.5	73.4	93.8	97.9	79.6	91.6	93.5	67.5	82.0	85.3
PCB [55]	42.1	74.8	88.2	80.6	96.2	98.6	82.6	92.7	95.2	66.4	82.5	87.1
PyrNet [56]	43.6	79.2	90.4	83.9	97.1	98.5	84.1	93.0	96.2	79.6	88.1	91.2
DDGAN [37]	51.2	83.6	94.4	85.7	97.1	98.6	-	-	-	-	-	-
JBIM + OSNet [39]	53.1	84.0	95.2	88.7	97.5	99.0	-	-	-	-	-	-
DGRL [42]	52.4	78.0	84.8	98.1	99.6	99.6	93.9	98.0	98.8	-	-	-
Ours-ResNet-50	**63.6**	79.2	**96.6**	89.2	98.9	**99.8**	90.1	96.2	97.7	81.9	**92.4**	**94.0**
Ours-Swin-B	**64.0**	**84.7**	**96.2**	**96.4**	**99.1**	**99.8**	**93.3**	**98.0**	**97.9**	**83.0**	**93.3**	**94.9**
Ours-Swin-L	**65.8**	**89.1**	**96.6**	**98.2**	**99.8**	**99.8**	**94.1**	**98.9**	**98.9**	**85.2**	**94.8**	**95.8**

TABLE II: Impact of the varied length of sub-vectors with respect to the rank-1 and mAP values on MLR-CUHK03. The query image is down-sampled at resolution $k = 1/2$, while the encoding length len is varied from 1 to 4. E2E and Prog stand for end-to-end and progressive training, respectively.

Model	$len = 1$		$len = \frac{1}{2}$		$len = \frac{1}{3}$		$len = \frac{1}{4}$	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
E2E	78.6	75.0	**87.8**	**87.0**	82.4	80.6	81.4	80.9
Prog	80.3	78.7	**89.2**	**89.0**	83.7	82.4	83.0	81.8

B serves as a viable alternative to Swin-L, especially when computational costs are a significant factor to consider. **(3)** Comparing with standard re-ID methods, such as CamStyle [53] and FD-GAN [54], which can also be applied in CRReID, our method (e.g., Ours-ResNet-50) still outperforms those competitors.

2) Length of Sub-vectors: In this experiment, we investigate the effect of the length of sub-vectors on the feature representation. More specifically, for a predefined resolution, i.e., $k = 1/2$, we vary the number of sub-vectors in the query image feature vector as such the length of sub-vectors is varying. As shown in Table II, our model encodes the query resolution into the feature vector with its length portion $len = \{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$, wherein len=1 indicates the full-dimension feature vector and $len = \frac{1}{k}$ indicates the proportionate length of the feature vector. Our model variations, i.e., end-to-end and progressive training, both reveal that encoding the query resolution $k = 1/2$ into the appropriate length of sub-vectors, i.e., $len = \frac{1}{2}$, enables the aligned comparison with the gallery image at full resolution because the shared information is reflected in the corresponding sub-vectors.

E. Ablation Studies

This section performs ablation studies to examine the effectiveness of resolution adaptive representations and the impact of various components of our method.

1) Comparison with A Naive Solution for CRReID: One naive solution to realize query-adaptive metric is to build k versions of gallery images, with each one corresponding to a possible level of image resolution. Based on the query image resolution (unseen query resolution could be assigned to one of the nearest resolutions), one can pick the corresponding version of gallery images to make comparison.

To handle the resolution mismatch, a trivial solution is to simply down-sample the HR images such that the resolution is compatible to the LR query. As such, it is intriguing to know whether it is necessary to perform resolution-adaptive representation learning. To verify this, we train several naive baselines on the three datasets. Specifically, for both MLR-CUHK03 and MLR-Market-1501, we first down-sampled the HR training images at different resolutions {1/2,1/3,1/4} to match the query at the corresponding resolution. Then, using the two losses (i.e., identity loss and verification loss), we trained a naive baseline with a group of LR query and LR gallery images under obtain a resolution specific matching model, e.g., $k = 1/2$ for each query-gallery pair. The matching results from different resolutions were averaged to form the reported final values. Since CAVIR is the only real cross-resolution dataset and its query presumably shows $k = 1/2$, we down-sampled the HR training images to form the LR counterpart. For a fair comparison, we also trained our model with the proposed two mechanisms in an end-to-end manner (ResNet-50 was used as backbone). All the comparison results are reported in Table III. This could be because our model can use the training samples from multiple-resolutions. We can see that the proposed model (trained in end-to-end or progressive) has an obvious advantage over the baseline on the three datasets. Interestingly, we find that the progressive training method consistently performs better than end-to-end training, with an improvement ranging from 2%-5%.

2) Study of Two Resolution-Adaptive Mechanisms: One may wonder the relative contribution of the proposed two resolution-adaptive mechanisms, i.e., a varying-length feature representation learning and learnable masks for intermediate

TABLE III: Comparison against a naive solution: training k retrieval models, one corresponding to a possible query resolution. Given a query image with resolution level k, its corresponding model is picked to perform retrieval. Our results show that our method outperforms this naive solution significantly. Note that all of our models use ResNet-50 as backbone.

Model	CAVIAR			MLR-CUHK03			MLR-Market-1501			MLR-DukeMTMC-reid		
	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
Naive baseline	18.1	44.8	56.4	83.3	96.0	97.1	85.6	94.2	96.0	76.8	90.2	93.1
Ours (end-to-end)	58.2	73.4	89.8	87.3	97.7	98.7	89.5	95.8	97.4	80.3	90.8	93.3
Ours (progressive)	**63.6**	**79.2**	**96.6**	**89.2**	**98.9**	**99.8**	**90.1**	**96.2**	**97.7**	**81.9**	**92.4**	**94.0**

TABLE IV: Investigation on different components of the proposed method (backbone on ResNet-50). IDE+Verif is the baseline with both identity classification loss and verification loss. Ours (w/o x) means removing the component x from our approach. Ours (w/o val)* denotes that our model with block-wise masks is trained in an end-to-end fashion. Also note that "val" denotes the varying-length representation.

Model	CAVIAR			MLR-CUHK03			MLR-Market-1501			MLR-DukeMTMC-reid		
	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
IDE + Verif	33.2	56.4	65.2	76.5	95.7	98.2	81.6	92.0	95.1	72.3	83.0	88.7
Ours (w/o mask)	57.4	69.5	75.6	85.0	97.2	98.6	85.1	94.2	96.5	78.4	88.6	91.6
Ours (w/o val)	57.2	71.6	77.2	85.8	97.4	98.7	89.3	95.8	97.4	80.7	91.2	94.0
Ours (w/o val)*	48.7	68.7	75.6	85.0	97.1	98.6	89.1	95.6	97.4	80.2	90.1	93.7
Ours (mask+val)	**63.6**	**79.2**	**96.6**	**89.2**	**98.9**	**99.8**	**90.1**	**96.2**	**97.7**	**81.9**	**92.4**	**94.0**

TABLE V: Study of loss functions on MLR-CUHK03. Best results are in boldface.

Method	MLR-CUHK03				
	Rank-1	Rank-5	Rank-10	Rank-20	mAP
Ours	**89.2**	**98.9**	**99.8**	**99.9**	**88.6**
Ours w/o $\mathcal{L}_{\text{cls}}$	77.4	94.9	97.4	98.6	76.8
Ours w/o $\mathcal{L}_{\text{verif}}$	76.3	94.5	97.0	98.5	75.8

activations, to the performance of CRReID. To study the impact of each component, we created four variants of our method: **(1) Ours (w/o mask)**, which removes the learnable masks and only uses the varying-length feature representations. **(2) Ours (w/o val)**, which does not use the varying-length feature representations but with the learnable masks. **(3) Ours (w/o val)***, which also only uses the learnable masks, but trains in an end-to-end fashion rather than adopting the progressive training strategy. **(4)** A baseline termed **IDE+Verif** was trained using the proposed framework without the learnable masks and varying-length representations. In IDE+Verif, a LR query is directly compared with a HR gallery without any resolution down/up-sampling. This baseline is trained in end-to-end. The obtained experimental results are shown in Table IV. We can see that each component, i.e., the learnable masks, varying-length feature representations and the progressive training, plays a critical role, and removing any of them will lead to degraded performance. Their combination leads to the best accuracy. Also, by comparing against the baseline approach, i.e., **IDE+Verif**, we observe that using either mechanism alone leads to a significant improvement. This illustrates the effectiveness of both mechanisms.

3) Study of Loss Functions: Our network is trained with two types of loss functions, i.e., the identification loss ($\mathcal{L}_{\text{cls}}$) and the verification loss ($\mathcal{L}_{\text{verif}}$). Thus, it is important to analyze the impact of different loss functions on network training. To this end, we ablate the two loss functions by comparing 2 variants of our model on the MLR-CUHK03 dataset. Table V reports the ablation study on the loss functions. When the loss $\mathcal{L}_{\text{cls}}$ is turned off, our method sees its rank-1 value drop from 89.2% (with two loss functions) to 77.4%. Without the loss of $\mathcal{L}_{\text{verif}}$, our model only achieves 76.3% at rank-1. This demonstrates that both loss functions are crucial in our method. This is consistent with observations in the existing literature [58]. This suggests that the $\mathcal{L}_{\text{cls}}$ plays an important role in cross-resolution re-ID by separating the identities in the feature space. Without loss $\mathcal{L}_{\text{verif}}$, our model only achieves 76.3% at rank-1. This demonstrates that the verification loss is also crucial to our model. The loss $\mathcal{L}_{\text{verif}}$ is able to regularize the feature embeddings of different resolutions with intra-class compactness. We conclude that the combination of two loss functions can achieve the superb evaluation results in terms of rank-1, -5, -10, -20 and mAP on MLR-CUHK03.

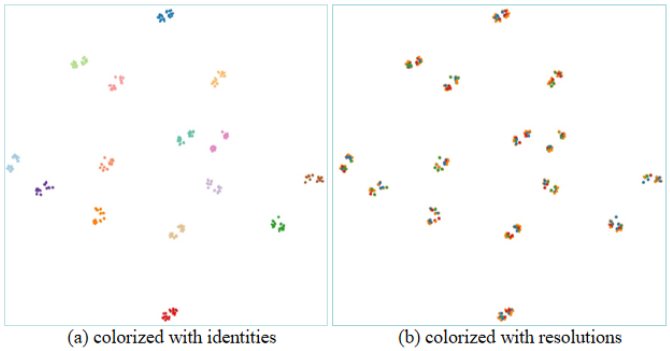
4) Top-ranked Gallery w.r.t Varied-resolution Queries: Given a query at different resolutions, we present the first top-15 ranked gallery images from MLR-Market-1501 in Fig. 4. The green and red rectangles indicate the correct and incorrect matches, respectively. The first row of Fig. 4 shows the ranking results using the query with its original resolution and matched against HR gallery. When the query has lower resolution, e.g., $k = 1/3$ or 1/4, corresponding to the third and fourth rows of Fig. 4, our method can still achieve 13 correct matches out of 15 candidate images for the low-quality queries. This demonstrates the effectiveness of our method in addressing the resolution mismatch between the query and the gallery.

5) Embedding of Query-Adaptive Features: To demonstrate the effectiveness of our method in deriving resolution-adaptive features for different resolutions, we visualize the feature vectors of images from the MLR-CUHK03 *test set* in Fig. 5. More specifically, we select 15 different identities, each of which is described by a specific color, and we project the feature vectors in 2D feature space using t-SNE. The projection is shown in Fig. 5 (a). We observe that our model can establish a well-separated feature space for re-ID. To close up the distribution of different resolutions, we colorize each resolution with a different color in each identity cluster, i.e., four different colors for four resolutions $r \in \{1, 2, 3, 4\}$, and project features vectors via t-SNE. The results are shown in Fig. 5 (b). Again we observe that the projected feature vectors of the same identity but different down-sampling rates are well separated. The visualization results demonstrate that our method learns resolution-adaptive representations that are effective for cross-resolution person re-ID.

Fig. 4: The top-15 ranked gallery images w.r.t the HR query and its down-sampled LR queries with down-sampling rates $r \in \{2, 3, 4\}$. The ranking evaluation is performed on MLR-Market-1501 dataset. The correct and incorrect matching gallery images are displayed in green and red rectangles, respectively.

TABLE VI: Study of ***unseen*** down-sampled resolutions on MLR-CUHK03. The values in brackets indicate the values obtained by turning the unseen resolution into *seen* resolution for training. $x^L_{1/3} \rightarrow x^L_{1/2}$ means assigning the unseen $x^L_{1/3}$ to the down-sampling rate $x^l_{1/4}$ which is seen during training. Note that "Prog" denotes progressive training and "E-E" denotes end-to-end. (Best view in color)

Test (***unseen***)		$x^L_{1/3} \rightarrow x^L_{1/2}$				$x^L_{1/3} \rightarrow x^L_{1/4}$			
		Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
$x^L_{1/3}$	Prog	83.4 (*85.6*)	97.2 (*97.3*)	98.6 (*98.6*)	99.3 (*99.6*)	83.0 (*85.6*)	96.9 (*97.3*)	98.6 (*98.6*)	99.4 (*99.6*)
	E-E	79.5 (*81.5*)	96.4 (*96.9*)	98.6 (*98.6*)	99.3 (*99.3*)	79.8 (*81.5*)	96.5 (*96.9*)	98.4 (*98.6*)	99.1 (*99.3*)
Test (***unseen***)		$x^L_{1/6} \rightarrow x^L_{1/4}$				$x^L_{1/6} \rightarrow x^L_{1/8}$			
		Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
$x^L_{1/6}$	Prog	80.5 (*84.1*)	96.4 (*97.2*)	98.6 (*98.6*)	99.4 (*99.5*)	81.0 (*84.1*)	96.8 (*97.2*)	98.8 (*98.6*)	99.6 (*99.5*)
	E-E	76.5 (*79.7*)	95.7 (*96.6*)	98.1 (*98.7*)	98.9 (*99.4*)	77.9 (*79.7*)	96.2 (*96.6*)	98.6 (*98.7*)	99.2 (*99.4*)



(a) colorized with identities (b) colorized with resolutions

Fig. 5: The t-SNE visualization of the learned resolution-adaptive features on the MLR-CUHK03 test split. (a) The embedding of identity features. Each color corresponds to one identity. (b) The same data which is colorized to show resolution specifics with four colors corresponding to four down-sampling rates. Best viewed in color.

F. Generalization to Unseen Resolutions

In the standard setting for cross-resolution person re-ID, the resolutions or down-sampling rates at the test time are seen during training. In practice, we may encounter the scenario that the test image has a resolution that is not seen during training. It is important that our algorithm could handle this case. To this end, we propose the following scheme: **(1)** we train our model with a set of fixed down-sampling rates, e.g., $r = \{2, 4, 6, 8\}$. **(2)** for a test image with unseen down-sampling rate, say $r = 3$, we simply assign the test image to the nearest down-sampling rate seen during training and process it as the assigned down-sampling rate. For example, if the resolution of a test image is equivalent to down-sampling the HR image 3 times, we treat the test image as if its ratio is 2 or 4 and run our algorithm. Note that $r = 8$ indicates a very low resolution.

To evaluate the effectiveness of the above scheme, we conduct the following experiment, in which we construct a new MLR dataset on CUHK03 by using the down-sampling rates $r \in \{2, 4, 8\}$, and then train the model with the progressive and end-to-end mask learning schemes. At the test stage, we consider the query with two unseen resolutions/down-sampling rates, i.e., $r = 3$ and $r = 6$, denoted as $x^L_{1/3}$ and $x^L_{1/6}$, respectively. Following the principle aforementioned, we can assign it to $r = 2, 4$ or $r = 4, 8$, respectively. We evaluate their performance against baselines that are trained with $r = 3$ and $r = 6$ images. The results are presented in Table VI. We can observe that assigning an unseen resolution to the nearest training resolution leads to reasonably good performance. In comparison to the baseline which makes unseen resolution seen during training (i.e., we train a new model with the down-sampling rates $r \in \{2, 3, 6\}$), the performance drop is around 2% for $k = 1/3$ and is around 4% for $k = 1/6$ in rank-1. The performance difference becomes much smaller from rank-5 to -20. This suggests that this proposed simple solution is sufficient to handle unseen resolutions at test time. Also, we observe that assigning the unseen resolution to its higher or lower resolution proxy does not make much difference. The

TABLE VII: Study of feature dimension division on MLR-CUHK03.

Method	MLR-CUHK03				
	Rank-1	Rank-5	Rank-10	Rank-20	mAP
ResNet-50	**89.2**	**98.9**	**99.8**	**99.9**	**88.6**
Gbumbel-Softmax	84.1	93.5	95.0	96.2	81.9

performance is largely comparable in most cases.

When the unseen resolution, e.g., $x^L_{1/3}$ is issued in test, we can approximate the feature vector for $x^L_{1/3}$ by using a nearby seen resolution, i.e., $x^L_{1/3} \rightarrow x^L_{1/2}$ and $x^L_{1/3} \rightarrow x^L_{1/4}$. We can see that approximating the unseen resolution using a lower down-sampling rate achieves better results than using a higher rate. For instance, in the case of matching a very low resolution query, i.e., $x^L_{1/6}$ using $x^L_{1/6} \rightarrow x^L_{1/8}$ outperforms using $x^L_{1/6} \rightarrow x^L_{1/4}$ in both continual and end-to-end training. In comparison, we turn the unseen resolutions into seen resolutions by re-training the model with down-sampling rates $r \in \{2, 3, 6\}$ and report the results in the brackets of Table VI. Our method is seen to achieve results similar to the case of training the model with a seen resolution, For example, for the unseen $x^L_{1/3} \rightarrow x^L_{1/2}$ it only drops the rank-1 by 2.2% compared with the training with the specific resolution (rank-1=83.4% v.s rank-1=85.6% for turning the unseen into a seen resolution in training).

G. Division on Sub-feature Dimension for Resolutions

Our method hypothesizes that the length of a feature vector corresponds to its resolution, which can be implemented by discretisizing the resolution. Thus, one may wonder if there is a viable approach to automate the feature vector division and allows for automating continuous sub-feature dimension. In this experiment, we employ the Gumbel-Softmax [59], where a resolution variable Z [2] is parameterised as the resolution distribution $\pi_1, \ldots, \pi_x$ and π_i is the resolution possibility to be learned by the neural network. Experimental results are reported in Table VII. One primary reason is that the re-parameterisation on the continuous sampling for Z may not effectively differentiate the resolution differences, leading to performance drop.

VI. Conclusion

In this paper, we present a novel approach to produce resolution-adaptive representations for cross-resolution person re-identification (CRReID). Specifically, we propose two novel adaptation mechanisms: a varying-length representation learning to produce the feature vector with varied dimensions corresponding to resolution levels, and a set of resolution-adaptive masks applied to intermediate feature blocks to further enhance the resolution disentanglement. The two strategies are slotted together to achieve the SOTA performance on multiple CRReID benchmarks, especially the merits of addressing the resolution mismatch issue. Future work could explore generative models [60] to reconstruct resolutions for the proposed method.

[2] Z is an one-hot vector that determines the resolution of an image.

Acknowledgement

This work was partially funded by Australian Research Council (Grants DP210101682 and DP210102674) and NSFC 62172136, U19A2073, 62002096.

References

[1] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by- tracking," in *CVPR*, 2008.

[2] R. Vezzani, D. Baltieri, and R. Cucchiara, "People re-identification in surveillance and forensics: A survey," *ACM Computing Surveys*, vol. 46, no. 2, pp. 1–37, 2013.

[3] M. Jia, X. Cheng, Y. Zhai, S. Lu, S. Ma, Y. Tian, and J. Zhang, "Matching on sets: Conquer occluded person re-identification without alignment," in *AAAI*, vol. 35, no. 2, 2021.

[4] L. Tan, P. Dai, R. Ji, and Y. Wu, "Dynamic prototype mask for occluded person re-identification of feature detectors," in *ACMMM*, 2022.

[5] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *AAAI*, 2018.

[6] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, X. Du, and Y.-C. F. Wang, "Recover and identify: A generative dual model for cross-resolution person re-identification," in *ICCV*, 2019, pp. 431–440.

[7] K. Han, Y. Huang, Z. C. L. Wang, and T. Tan, "Prediction and recovery for adaptive low-resolution person re-identification," in *ECCV*, 2020.

[8] G. Zhang, Y. Chen, W. Lin, A. Chandran, and X. Jing, "Low resolution information also matters: Learning multi-resolution representations for person re-identification," in *IJCAI*, 2021, pp. 1295–1301.

[9] G. Zhang, G. Yu, Z. Dong, H. Wang, Y. Zheng, and S. Chen, "Deep high-resolution representation learning for cross-resolution person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 8913–8925, 2021.

[10] Y.-C. Chen, Y.-J. Li, X. Du, and Y.-C. F. Wang, "Learning resolution-invariant deep representations for person re-identification," in *AAAI*, 2019.

[11] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," in *arXiv:1207.0580v1*, 2012.

[12] L. Wu, Y. Wang, J. Gao, M. Wang, Z.-J. Zha, and D. Tao, "Deep co-attention based comparators for relative representation learning in person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 722–735, 2021.

[13] L. Wu, R. Hong, Y. Wang, and M. Wang, "Cross-entropy adversarial view adaptation for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2081–2020, 2020.

[14] L. Wu, Y. Wang, X. Li, and J. Gao, "What-and-where to match: Deep spatially multiplicative integration networks for person re-identification," *Pattern Recognition*, vol. 76, pp. 727–738, 2018.

[15] L. Wu, Y. Wang, H. Yin, M. Wang, and L. Shao, "Few-shot deep adversarial learning for video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 1233–1245, 2020.

[16] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *ECCV*, 2018.

[17] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017, pp. 1077–1085.

[18] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017, pp. 3960–3969.

[19] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *ECCV*, 2018.

[20] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, "Beyond human parts: Dual part-aligned representations for person re-identification," in *ICCV*, 2019, pp. 3642–3651.

[21] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *CVPR*, 2018, pp. 1179–1188.

[22] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *ACM Multimedia*, 2018.

[23] Z. Wei, X. Yang, N. Wang, and X. Gao, "Flexible body partition-based adversarial learning for visible infrared person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4676 – 4687, 2022.

[24] ——, "Syncretic modality collaborative learning for visible infrared person re-identification," in *ICCV*, 2021, pp. 225–234.

[25] L. Wu, D. Liu, W. Zhang, D. Chen, Z. Ge, F. Boussaid, M. Bennamoun, and J. Shen, "Pseudo-pair based self-similarity learning for unsupervised person re-identification," *IEEE Transactions on Image Processing*, vol. 831, pp. 4803–4816, 2022.

[26] R. Zhou, X. Chang, L. Shi, Y.-D. Shen, Y. Yang, and F. Nie, "Person re-identification via multi-feature fusion with adaptive graph learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, pp. 1592–1601, 2019.

[27] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *ICCV*, 2019.

[28] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, 2018, pp. 2285–2294.

[29] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *ICCV*, 2017, pp. 350–359.

[30] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *CVPR*, 2018, pp. 994–1003.

[31] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *CVPR*, 2019, pp. 98–607.

[32] K. Zheng, W. Liu, L. He, T. Mei, J. Luo, and Z.-J. Zha, "Group-aware label transfer for domain adaptive person re-identification," in *CVPR*, 2020, pp. 5310–5319.

[33] Z. Bai, Z. Wang, J. Wang, D. Hu, and E. Ding, "Unsupervised multi-source domain adaptation for person re-identification," in *CVPR*, 2021, pp. 12 914–12 923.

[34] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. long Liu, D. Yue, R. Hu, and B. Xu, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *CVPR*, 2015.

[35] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *ICCV*, 2015, pp. 3765–3773.

[36] Z. Wang, R. Hu, Y. Yu, J. Jiang, C. Liang, and J. Wang, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *IJCAI*, 2016.

[37] Y. Huang, Z.-J. Zha, X. Fu, R. Hong, and L. Li, "Real-world person re-identification via degradation invariance learning," in *CVPR*, 2020, pp. 14 084–14 094.

[38] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, and Y.-C. F. Wang, "Cross-resolution adversarial dual network for person re-identification and beyond," in *arXiv:2002.09274v2*, 2020.

[39] W.-S. Zheng, J. Hong, J. Jiao, A. Wu, X. Zhu, S. Gong, J. Qin, and J. Lain, "Joint bilateral-resolution identity modeling for cross-resolution person re-identification," *International Journal of Computer Vision*, vol. 130, pp. 136–156, 2022.

[40] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh, "Cascaded sr-gan for scale-adaptive low resolution person re-identification," in *IJCAI*, 2018.

[41] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, and Z. Wang, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017, pp. 4681–4690.

[42] X. Ye and G. Gao, "Cross-resolution person re-identification via deep group-aware representation learning," in *ICPR*, 2022, pp. 863–869.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[44] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *CVPR Workshop*, 2019, pp. 4321–4329.

[45] L. Yang, Z. He, J. Zhang, and D. Fan, "Fast multiple task adaptation via kernel-wise soft mask learning," in *CVPR*, 2021, pp. 13 845–13 853.

[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[48] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *BMVC*, 2011.

[49] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deep reid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2011.

[50] L. Zheng, L. Shen, L. Tian, S. Wang, J. dong Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[51] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV Workshop on Benchmarking Multi-Target Tracking*, 2016.

[52] Z. Cheng, Q. Dong, S. Gong, and X. Zhu, "Inter-task association critic for cross-resolution person re-identification," in *CVPR*, 2020, pp. 2605–2615.

[53] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *CVPR*, 2018.

[54] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, and H. Li, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," 2018.

[55] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *ECCV*, 2018.

[56] N. Martinel, G. L. Foresti, and C. Micheloni, "Aggregating deep pyramidal representations for person re-identification," in *CVPR Workshop*, 2019.

[57] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017, pp. 1318–1327.

[58] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *IJCAI*, 2017, pp. 2194–2200.

[59] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *ICLR*, 2017, pp. –.

[60] C. Yan, X. Chang, Z. Li, W. Guan, Z. Ge, L. Zhu, and Q. Zheng, "Zeronas: Differentiable generative adversarial networks search for zero-shot learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 44, pp. 9733–9740, 2021.