

# Person Re-Identification via Generalized Class Prototypes

Md Ahmed Al Muzaddid<sup>✉</sup> and William J. Beksi<sup>✉</sup>

The University of Texas at Arlington, Arlington TX 76019, USA

**Abstract.** Advanced feature extraction methods have significantly contributed to enhancing the task of person re-identification. In addition, modifications to objective functions have been developed to further improve performance. Nonetheless, selecting better class representatives is an underexplored area of research that can also lead to advancements in re-identification performance. Although past works have experimented with using the centroid of a gallery image class during training, only a few have investigated alternative representations during the retrieval stage. In this paper, we demonstrate that these prior techniques yield suboptimal results in terms of re-identification metrics. To address the re-identification problem, we propose a generalized selection method that involves choosing representations that are not limited to class centroids. Our approach strikes a balance between accuracy and mean average precision, leading to improvements beyond the state of the art. For example, the actual number of representations per class can be adjusted to meet specific application requirements. We apply our methodology on top of multiple re-identification embeddings, and in all cases it substantially improves upon contemporary results.

**Keywords:** Image Retrieval · Person Re-Identification

## 1 Introduction

Person re-identification (Re-ID) has been widely studied as a specific image retrieval problem. Given a query image of a person, the goal is to identify it from a set of gallery images captured by a group of non-overlapping cameras. The gallery images typically include disjoint views of the same person taken by different cameras at distinct times. Re-ID research has undergone rapid growth since the introduction of initial datasets [6,8] for this task. The technology has found widespread application in various domains such as autonomous vehicles, security and surveillance systems, sports analytics, and much more. The Re-ID task is challenging due to a combination of dynamic lighting conditions, low image resolution, multiple camera viewpoints, occlusions, unconstrained poses, and unreliable bounding boxes.

A common approach to person Re-ID is to first transform the query and gallery images into feature vectors using either handcrafted feature engineering or deep learning feature extractors. Subsequently, the similarity between the query and gallery feature vectors is assessed by measuring the distance between them. The greater the proximity of these vectors, the higher the degree of similarity. More specifically, we expect that the feature vectors representing the same person will exhibit a close spatial relationship,

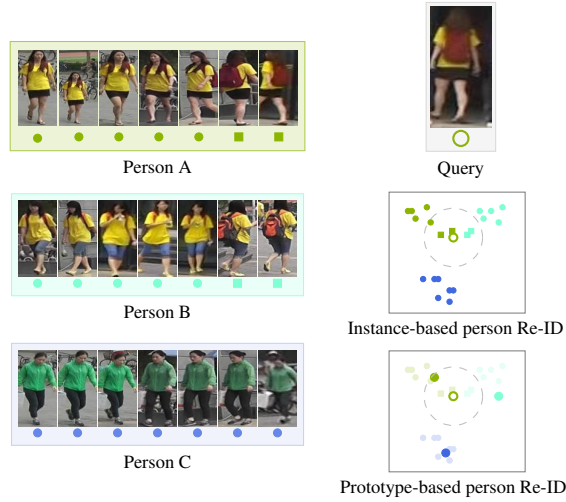


Fig. 1: An overview of prototype-based Re-ID. The solid circles/squares denote the feature vectors of each image instance, while the empty circles represent query vectors. The two plots on the right depict the feature vector distribution in the embedding space. The upper-right plot illustrates instance-based Re-ID, which typically yields lower precision. The plot depicts a specific case where among the five nearest instances from the query, two are false positives. The lower-right plot demonstrates how our prototype-based Re-ID approach enhances precision by representing all instances from the same class using a distribution-aware feature vector.

while vectors representing different individuals will have a significant separation in the vector space where distance is measured by the Euclidean norm.

Since images for person Re-ID are typically transformed into feature vectors, feature extraction via deep learning has become a key aspect in recent advancements. This research has focused on developing more discriminative features, yet it has inherent limitations. Specifically, it is restricted to extracting viewpoint-centric features from a single image, which restricts its ability to form a comprehensive, person-centric representation by utilizing multiple images of the same individual. Furthermore, the extensive image comparisons required for Re-ID impose high computational demands, restricting its practical application in resource-constrained environments.

As an alternative approach, Snell et al. [23] introduced the prototypical network for Re-ID, a method that learns a metric space in which classification is performed by calculating distances to prototype representations for each class. Building on this paradigm, we propose a generalization of this concept by employing multiple prototypes per class. We show that using multiple prototypes per class enhances the performance of Re-ID tasks compared to a single prototype per class. *To the best of our knowledge, we are the first to establish the effectiveness of varying numbers of prototypes for image retrieval.*

Not only do we examine prototypes as class representatives, but we also generalize the class representative selection process. In particular, we develop a transformer decoder-based model that takes a set of images of an object of interest and generates one or more set/class representatives in the embedding space during inference. When

generating multiple prototypes iteratively, the decoder’s self-attention module attends to all prototypes generated in preceding iterations, while the cross-attention module is conditioned on the overall feature distributions of the entire class. Fig. 1 illustrates the selection of class representatives via our method compared to conventional (e.g., instance-based) approaches.

To demonstrate the effectiveness of our approach, we present baseline algorithms for efficiently identifying robust prototypes. These methods sample class representatives that capture intra-class diversity while maintaining inter-class boundaries within the embedding space. In summary, our contributions are as follows.

- We provide an analysis of the impact of different class representatives (e.g., instance, centroid) on person Re-ID tasks.
- We create an attention-based model to generate class prototypes that can be directly compared with the query during inference time.
- We present multiple sampling-based algorithms for selecting class representatives and establish state-of-the-art person Re-ID benchmarks.

Our source code is publicly available at [7].

## 2 Related Work

### 2.1 Prototype-Based Methods

Similar to the prototypical network approach introduced by Snell et al. [23], Wang et al. [30] developed embedding models that classify each query example by calculating distances to speaker prototypes represented by centroids. Other research has explored alternative strategies for aggregating image features. For example, centroid vectors have been utilized during model training across various applications [12,34,37].

Recently, a centroid-based approach was used to represent gallery classes during inference for person Re-ID [33]. This improves performance by summarizing all gallery images of a single class into an average centroid vector, enabling similarity measurement through distances between query features and representations. However, we do not rely on a single fixed prototype such as a centroid. Instead, our approach allows flexibility in selecting a predefined number of representative points within the embedding space based on the feature distribution.

### 2.2 Attention Models

Attention models have become very effective for person Re-ID. They can handle the part alignment challenge and enhance feature representation. For instance, Liu et al. [17] proposed an end-to-end comparative attention network that learns to focus on part pairs of images of people to compare their appearance. Researchers have also employed attention models over a sequence of frames to extract salient features [18,14,22]. More recently, Zhang et al. [36] developed a patch-wise augmentation technique to enhance the representation of high-frequency components in attention-based models.

In this work, given a set of samples with their corresponding camera ID, we use an attention model to find prototypes that represent a class. Our approach exhibits parallels

with prior works such as [40,9,29], in which the inference time setting allows the model to incorporate group information. Nevertheless, these past works modify individual instances based on the group information while we utilize this information to provide a reduced number of new prototypes.

### 2.3 Loss Functions

A substantial body of research delves into various loss functions for Re-ID. The center loss, developed by Wen et al. [32], facilitates concurrent learning of feature centers for each class and effectively constrains large distances between features and their respective class centers. A quadruplet loss function able to model an output with a larger inter-class variation and a smaller intra-class variation, when compared to the triplet loss, was presented by Chen et al. [2]. Additionally, there is the pair-wise contrastive [4] and triplet ranking [41] losses. Zhu et al. [44] focused on the heterogeneity of the data and developed a hetero-center loss to reduce intra-class cross-modality variations. In contrast to dense comparisons that use only a select number of suitable pairs for each class within a mini-batch, a sparse pair-wise loss method was proposed by Zhou et al. [43]. We implement a custom loss function derived from the triplet and contrastive losses to enforce larger inter-class and smaller intra-class distances among the prototypes.

## 3 Method

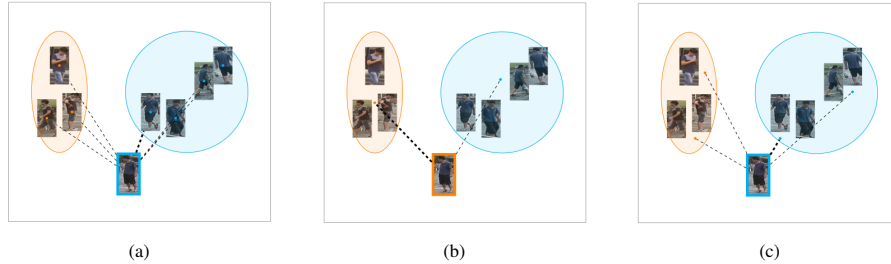


Fig. 2: A comparison among (a) instance-based, (b) centroid-based, and (c) prototype-based Re-ID. The images enclosed by the colored rectangles are the query images and the small colored dots are the class representatives. The dashed lines indicate the distance between the query image and the class representation. In (a), (b), and (c), the query image is assigned to the class of the nearest instance, centroid, and prototype (our method) respectively.

### 3.1 Motivation

We first motivate our work by analyzing the decision boundaries and hypothesis spaces associated with instance-based and centroid-based Re-ID methods. We begin with the instance-based method, a case where each gallery image is represented by its feature vector (Fig. 2a). In this approach, the class label for a query image is assigned based on the most similar gallery image, with similarity measured by calculating the distance

between the query image’s feature vector and those of all the gallery images. For the instance-based method, the decision boundary can be described in terms of the gallery instances,

$$\|x^{c_i} - b\| = \|x^{c_j} - b\|, \quad (1)$$

where  $b$  is a point on the decision boundary, and  $x^{c_i}$  and  $x^{c_j}$  are the two closest gallery instances from  $b$ , representing two distinct classes (i.e.,  $c_i$  and  $c_j$ ). This decision boundary can become quite complex depending on the distribution of the gallery instances.

The associated hypothesis space has a Vapnik-Chervonenkis (VC) dimension [26] of  $\mathcal{O}(n)$ , where  $n = |G|$  is the cardinality of the gallery set. In Re-ID tasks, where  $n \gg 100$ , the VC dimension of the instance-based approach increases substantially with the size of the gallery set. This expansion of the hypothesis space makes the decision boundary more susceptible to overfitting, which often results in a reduction of mean precision.

Conversely, the centroid-based Re-ID approach (Fig. 2b) imposes a highly-constrained hypothesis space that is restricted to only linear class boundaries of the form  $w \cdot x + b = 0$ , where  $x$  represents the centroid, and  $w$  and  $b$  are the coefficients defining the separating hyperplane. Due to the limited expressive capacity of this hypothesis space, it often underfits the gallery instances. Thus, an intermediate approach that strikes a balance between the overfitting tendencies of instance-based methods and the underfitting limitations of centroid-based techniques is needed.

### 3.2 Generalized Class Prototype

Image similarity is typically measured by calculating the distance between a query and a class representative (e.g., centroid) within a feature space. We refer to the class representative as a *class prototype* and define the retrieval process based on these prototypes as *prototype-based image retrieval*. Concretely, a class prototype can be any vector within the feature space for which there exists a function,

$$\mathcal{D}(p, q) = d \in \mathbb{R}, \quad (2)$$

that measures the distance  $d$  between the query  $q \in \mathbb{R}^n$  and the prototype  $p \in \mathbb{R}^n$ . The number of prototypes,  $N$ , can exceed one per class and may vary across different classes. We use the terms *class*, *individual*, and *identity* interchangeably.

Using these concepts, we define *generalized class prototype (GCP)* Re-ID as an image retrieval method in which a query image is re-identified by measuring its similarity to class prototypes (Fig. 2c). The centroid, which is the mean of all the gallery images of a class, is a special prototype. The Re-ID approach that quantifies similarity by measuring the distance between the query and the centroid is a specific instance of prototype-based image retrieval, where the number of prototypes per class is fixed (i.e.,  $N = 1$ ). We refer to this as the *centroid prototype*.

Traditional instance-based comparison, where each gallery image is individually compared with a query image, is also a special case of prototype-based retrieval but with a variable number of prototypes. More formally, the number of prototypes per class is denoted as  $N^{c_1}, N^{c_2}, \dots, N^{c_m}$  where  $N^{c_1} = |G_{c_1}|, N^{c_2} = |G_{c_2}|, \dots, N^{c_m} = |G_{c_m}|$ .

$G_{c_i}$  represents the set of gallery images belonging to class  $c_i$  and  $m$  is the total number classes in the gallery set. We refer to this type of prototype as the *instance prototype*.

Given the special cases of prototype-based Re-ID, we can now address the solution to the Re-ID problem in terms of a GCP. The objective is to assign prototypes  $p_k^{c_i}$  for any given query  $q^{c_i}$  associated with the true identity  $c_i$  such that the following condition holds,

$$\exists_{k \in \{1, \dots, N^{c_i}\}}, \forall_{c_j: c_j \neq c_i} \|p_k^{c_i} - q^{c_i}\| < \|p_{k'}^{c_j} - q^{c_i}\|, \quad (3)$$

where  $p_{k'}^{c_j}$  represents any prototype for identity  $c_j$ . Without an exact query in advance, we assume that the gallery and query images are independent and identically distributed. Under this assumption, gallery images can serve as proxies for the query images. This allows us to substitute  $q^{c_i}$  with  $x^{c_i}$  in (3) resulting in

$$\exists_{k \in \{1, \dots, N^{c_i}\}}, \forall_{c_j: c_j \neq c_i} \|p_k^{c_i} - x^{c_i}\| < \|p_{k'}^{c_j} - x^{c_i}\|. \quad (4)$$

A trivial solution to this inequality emerges if we assign  $p_k^{c_i} = x^{c_i}$ . This is similar to the instance-based Re-ID scenario where the number of prototypes per class is set to  $N^{c_1} = |G_{c_1}|, N^{c_2} = |G_{c_2}|, \dots, N^{c_m} = |G_{c_m}|$ . Restricting the number of prototypes per class to  $N^{c_i} < |G_{c_i}|$  makes solving (4) a non-trivial task. With a general gallery distribution and for any  $N^{c_i} < |G_{c_i}|$ , an exact solution may not always be achievable.

To solve this problem, we optimize the prototype selection process. Rather than seeking an exact solution, our goal is to minimize Re-ID errors. We employ an attention-based model that effectively captures the structure of the gallery distribution, enabling the generation of a set of prototypes per class. The number of prototypes per class is treated as a tunable hyperparameter. To reduce the number of hyperparameters, we constrain the model to generate a fixed number of prototypes for all classes, i.e.,  $\forall_i N^{c_i} = N$ . For any class where  $|G_{c_i}| < N$ , the number of prototypes can be adjusted by setting  $N^{c_i} = \min(N, |G_{c_i}|)$ .

### 3.3 Attention-Based Model

To acquire a comprehensive representation of an individual, it is essential to integrate subtle appearance cues observed across different images. However, not all features are equally important. The feature integration process should emphasize distinguishable human features while minimizing the influence of insignificant details. To achieve this objective, we introduce a GCP model that leverages the attention mechanism of a transformer decoder [27]. Our decoder-only GCP architecture is illustrated in Fig. 3 and described as follows.

Given an image  $x \in \mathbb{R}^{H \times W \times C}$  where  $H$ ,  $W$ , and  $C$  represent the height, width, and the number of channels, respectively, we first extract the feature vector  $\mathcal{F}(x): \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^D$  using a pretrained backbone model  $\mathcal{F}$ . A set of feature vectors  $\{\mathcal{F}(x_1^c), \mathcal{F}(x_2^c), \dots, \mathcal{F}(x_s^c)\}$  of class  $c$  are combined to encode distinct aspects of varying appearance. Camera IDs (e.g., 1, 2, 3, etc.) associated with each image are used to generate positional embeddings  $\psi \in \mathbb{R}^{S \times D}$ , introducing camera and pose information into the model. The feature vectors corresponding to different images are treated as tokens. A variable number ( $s \leq |G_c|$ ) of tokens ordered arbitrarily, representing the same class  $c$  (i.e., individual),

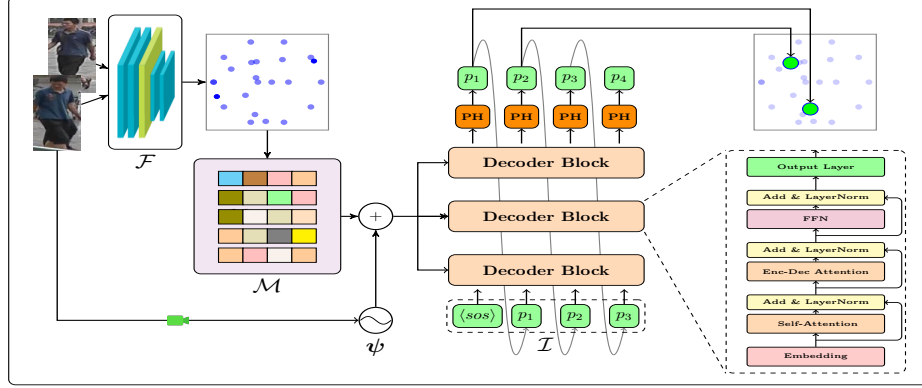


Fig. 3: The proposed attention-based GCP model for person Re-ID. Small blue dots represent the extracted features  $\mathcal{F}(x)$  from the backbone, while large green dots indicate the output prototypes from the model. PH denotes the prototype heads. For clarity, the figure displays only two of the generated prototypes.

are used as the memory of the decoder. Concretely, the memory fed into the GCP for class  $c$  is defined as

$$\mathcal{M}^c = [\mathcal{F}(x_1^c), \mathcal{F}(x_2^c), \dots, \mathcal{F}(x_s^c)] + \psi. \quad (5)$$

With  $\mathcal{M}^c$ , the GCP model generates  $N$  prototypes for class  $c$  in an autoregressive manner. We begin with an initial input sequence  $\mathcal{I}_1 = [\langle \text{sos} \rangle]$  of length one, containing a learned start-of-sequence token  $\langle \text{sos} \rangle$ . At each subsequent iteration  $t$ , we input the sequence  $\mathcal{I}_t = [\langle \text{sos} \rangle; p_1; p_2; \dots; p_{t-1}]$ , which includes the  $\langle \text{sos} \rangle$  token and all previously generated tokens  $[p_1; p_2; \dots; p_{t-1}]$ , into the decoder. We retain only the latest generated token  $p_t \in \mathbb{R}^D$  as the prototype for iteration  $t$ . Our auto-regressive approach is essential for creating prototypes step-by-step. Each generated prototype influences subsequent ones, which enables the model to produce prototypes that cover different regions within the embedding space.

Using the memory  $\mathcal{M}$  and input sequence  $\mathcal{I}$ , the GCP model is trained in mini-batches to minimize the distance between the prototypes  $p^c$  and feature vectors  $x^c$  of the same class  $c$ , while simultaneously maximizing the distance between the prototypes and feature vectors of different classes. We realize this objective by using following loss function:

$$\mathcal{L} = \mathcal{L}_{\text{triplet}} + \lambda \mathcal{L}_{\text{reg}}, \quad (6)$$

$$\mathcal{L}_{\text{triplet}} = \max(0, m + \|a - p\|_2 - \|a - n\|_2), \quad (7)$$

where  $m$  is the margin, and  $a$ ,  $p$ , and  $n$  are the anchor, positive, and negative feature vectors, respectively.  $\lambda$  serves as the constant regularization factor. When only the triplet loss  $\mathcal{L}_{\text{triplet}}$  is used as the loss function, prototypes of the same class (i.e.,

$p_1^c, p_2^c, \dots, p_N^c$ ) collapse to a single point in the feature space. To prevent this and encourage diversity, we introduce

$$\mathcal{L}_{reg} = \mathbb{1}\{k = k'\} \|p_k^c - p_{k'}^c\|_2 + \mathbb{1}\{k \neq k'\} \max(0, m - \|p_k^c - p_{k'}^c\|_2). \quad (8)$$

## 4 Evaluation

### 4.1 Datasets

A comprehensive evaluation was performed on three commonly used person Re-ID benchmarks: CUHK03-NP (labeled) [15], Market-1501 [39], and MSMT17 [31]. Table 1 summarizes the statistics of each dataset. The CUHK03-NP dataset presents a challenging Re-ID environment due to varying camera settings, which result in photometric transformations. Collected in front of a supermarket, the Market-1501 dataset uses six cameras (five high-resolution and one low-resolution). MSMT17, the largest dataset in the evaluation, was derived from 180 hours of video footage.

| Dataset     | Images  | Cameras | Train ID | Test ID |
|-------------|---------|---------|----------|---------|
| CUHK03-NP   | 13,164  | 2       | 1,160    | 100     |
| Market-1501 | 32,668  | 6       | 751      | 750     |
| MSMT17      | 126,441 | 15      | 1,041    | 3,060   |

Table 1: A summary of the dataset statistics.

### 4.2 Implementation Details

To demonstrate the robustness of our GCP model, we employed multiple backbones to extract raw feature vectors. Akin to Luo et al. [19], we used a ResNet50 backbone with several optimizations to obtain global features. Additionally, we conducted experiments using feature vectors extracted from TransReID [9] and PHA [36]. We extracted features from the model pretrained on the training split of the dataset. These extracted features serve as memory  $\mathcal{M}$  for the GCP model from which the final class prototype vectors  $p_i$  are derived (Fig. 3).

GCP comprises six individual decoder blocks, each with multi-headed self-attention to process inputs and multi-headed cross-attention to attend to memory based on the input. Each layer includes four attention heads, with a feed-forward network dimension of 512 and a dropout rate set to 0.2. Prototype heads PH, positioned above the decoder, share the same parameters. A PH consists of a single perceptron layer ( $\mathbb{R}^n \rightarrow \mathbb{R}^n$ ), where  $n$  represents the feature dimension. To identify the optimal number of prototypes for the GCP model, we empirically evaluated different numbers ( $N$ ) of prototypes. Although we generate  $N$  prototypes (tokens) at the output layer, our model is adaptable to any number of prototypes due to its autoregressive capability.

The GCP model was trained end-to-end using the loss function  $\mathcal{L}$  defined by (6). In the experiments,  $m$  was set to 1.2 and the network was trained using stochastic gradient

descent with a learning rate of 0.01, momentum equal to 0.9, and a weight decay of  $5 \times 10^{-4}$ . We used a feature dimension size of 2048 for the ResNet50 backbone and 3840 for TransReID and PHA backbones with a batch size of 128 (16 classes with 8 instances per class). In addition, the training images were augmented with horizontal flip and normalization.

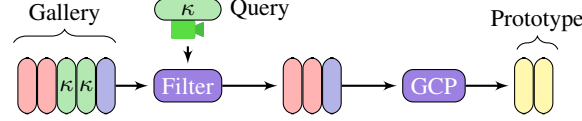


Fig. 4: Prototype generation during inference via the proposed GCP model. Each ellipse signifies a feature vector, with colors (e.g., red, green, blue) indicating the associated camera ID. When dealing with a particular query captured by a green camera, feature vectors marked in green are excluded during prototype generation.

During the inference stage, given a query  $q$  with class  $c$  and camera  $\kappa$ , we excluded all gallery instances from the same  $c$  that were captured using the same  $\kappa$  while generating gallery prototypes for  $c$ . This process is illustrated in Fig. 4. For prototypes of other classes, we used all available gallery features regardless of their cameras. We generated  $N$  prototypes sequentially in  $N$  iterations. The model was trained for 120 epochs on an Ubuntu 18.04 LTS machine with an Intel i7-8700 CPU, 64 GB of RAM, and an NVIDIA A100 GPU.

### 4.3 Experiments

The performance of our approach was measured using two metrics: cumulative matching characteristic at rank-1 (R-1) and mean average precision (mAP). Evaluations were conducted in a single-query setting without re-ranking. The GCP model demonstrated significant improvement in R-1/mAP across all base feature extractor models. Notably, with features extracted from the PHA model, the GCP model achieved 93.1%/92.2%, 97.3%/97.1%, and 89.7%/86.5% in R-1/mAP on the CUHK03-NP, Market-1501, and MSMT17 datasets, respectively.

We also compared against two baseline methods capable of generating  $N$  prototypes per class from the gallery set. The first and simpler approach is a clustering-based prototype selection algorithm, referred to as *k-centroid*. In this method, we applied k-means clustering to identify  $N$  clusters within each class and selected the cluster centroids as representative prototypes. The second method is based on farthest-point sampling (FPS) [5, 11]. FPS involves iteratively selecting points that are maximally distant from previously selected ones, thereby producing a subset that effectively approximates the diversity of the original distribution with fewer points.

In FPS, each feature vector  $\mathcal{F}(x) \in R^n$  is treated as a point. However, the standard FPS technique may select prototypes located at the class boundaries resulting in a poor mAP. To address this issue, we created a modified version called  $\alpha$ -farthest point sampling ( $\alpha$ -FPS), which is outlined in Algorithm 1. Additional details on the  $\alpha$ -FPS algorithm are provided in the appendix.

**Algorithm 1**  $\alpha$ -Farthest Point Sampling**Input:**  $X, N, \alpha$ **Output:**  $P$ 

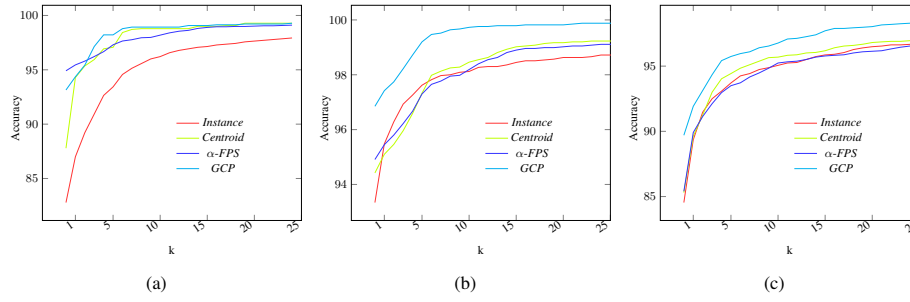

---

```

1:  $P \leftarrow \{\text{centroid}(X)\}$ 
2: while  $N \geq 1$  do
3:    $x, p \leftarrow \text{farthest}(X, P)$ 
4:    $X \leftarrow X \setminus x$ 
5:    $x \leftarrow x + \alpha(p - x)$ 
6:    $P \leftarrow P \cup \{x\}$ 
7:    $N \leftarrow N - 1$ 
8: end while

```

---

Fig. 5: Top- $k$  accuracy on the (a) CUHK03-NP, (b) Market-1501, and (c) MSMT17 datasets.

**Quantitative evaluation.** Table 2 provides a comparison of our results with both state-of-the-art Re-ID methods and baseline methods. We experimented with different values of  $N$  for the baseline methods and reported the best-performing results. On mAP, the GCP model surpasses most existing person Re-ID algorithms across all datasets. Notably, the 3840-dimensional feature vector extracted from the TransReID and PHA backbones exhibited superior performance compared to the 2048-dimensional feature vector generated by the ResNet50 backbone.

To ensure a fair comparison, we evaluated the GCP model alongside other prototype-based methods, including our baseline methods. We calculated top- $k$  accuracy for  $k \in \{1, 2, \dots, 25\}$ , with the results displayed in Fig. 5. These results show that the GCP approach consistently outperforms other methods in top- $k$  accuracy across all  $k$  values.

The number of images per person available in the gallery set varies as depicted in Fig. 6. Since GCP derives prototypes by aggregating features from multiple images of the same class, the number of images used in the prototype formation is expected to impact performance substantially. To assess this effect, we segmented the gallery set into distinct groups based on the number of images per class and evaluated model performance within each group. The results presented in Table 3 demonstrate that optimal performance is attained with 8 gallery images for the CUHK03-NP dataset, 10–20 gallery images for the Market-1501 dataset, and 16–30 gallery images for the MSMT17 dataset. Given the greater diversity and scale of the MSMT17 dataset, the model re-

| Method                              | CUHK03-NP   |             | Market-1501 |             | MSMT17      |             |
|-------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                                     | R-1         | mAP         | R-1         | mAP         | R-1         | mAP         |
| OSNet (ICCV'19)[42]                 | -           | -           | 94.8        | 84.9        | 78.7        | 52.9        |
| Pyramid (CVPR'19)[38]               | 78.9        | 76.9        | 95.7        | 88.2        | -           | -           |
| ABDNet (CVPR'19) [1]                | -           | -           | 88.3        | 95.6        | 60.8        | 82.3        |
| CBDB-Net (TCSVT'21)[25]             | 77.8        | 76.6        | 94.4        | 85.0        | -           | -           |
| Auto-ReID (CVPR'19) [21]            | 77.9        | 73.0        | 95.7        | 88.2        | -           | -           |
| st-ReID (AAAI'19)[28]               | -           | -           | 94.5        | 85.1        | -           | -           |
| C2F (CVPR'21)[35]                   | 80.6        | 79.3        | 94.8        | 87.7        | -           | -           |
| CDNet (CVPR'21) [13]                | -           | -           | 94.8        | 87.7        | 78.9        | 54.7        |
| PAT (CVPR'21)[16]                   | -           | -           | 95.4        | 88.0        | -           | -           |
| NFormer (CVPR'22)[29]               | 78.0        | 77.2        | 94.7        | 91.1        | 77.3        | 59.8        |
| BPBReID <sub>HR</sub> (WACV'23)[24] | -           | -           | 95.7        | 93.0        | -           | -           |
| SOLIDER (CVPR'23)[3]                | -           | -           | 95.7        | 89.4        | <b>90.7</b> | 77.1        |
| TransReID (CVPR'21)[9]              | 81.7        | 79.6        | 88.9        | 95.2        | 85.3        | 67.4        |
| SP loss (CVPR'23)[43]               | 82.4        | 84.6        | 89.6        | 80.5        | 82.3        | 61.0        |
| PHA (CVPR'23)[36]                   | 84.5        | 83.0        | 96.1        | 90.2        | 86.1        | 68.9        |
| IRM (CVPR'24)[10]                   | 86.5        | 85.4        | 96.5        | 93.5        | 86.9        | 72.4        |
| k-centroid + (PHA)                  | 85.9        | 84.2        | 94.4        | 94.8        | 85.1        | 81.1        |
| $\alpha$ -FPS + (PHA)               | 86.8        | 84.8        | 95.9        | 95.8        | 85.4        | 81.5        |
| GCP (ours) + (ResNet50)             | 88.6        | 88.1        | 89.5        | 95.4        | -           | -           |
| GCP (ours) + (TransReID)            | -           | -           | 95.3        | 95.9        | 85.4        | 82.1        |
| GCP (ours) + (PHA)                  | <b>93.1</b> | <b>92.2</b> | <b>97.3</b> | <b>97.1</b> | 89.7        | <b>83.4</b> |

Table 2: Quantitative results on the CUHK03-NP, Market-1501, and MSMT17 datasets. R-1 is top-1 accuracy and mAP is mean average precision. The best performance value for each column is marked in bold.

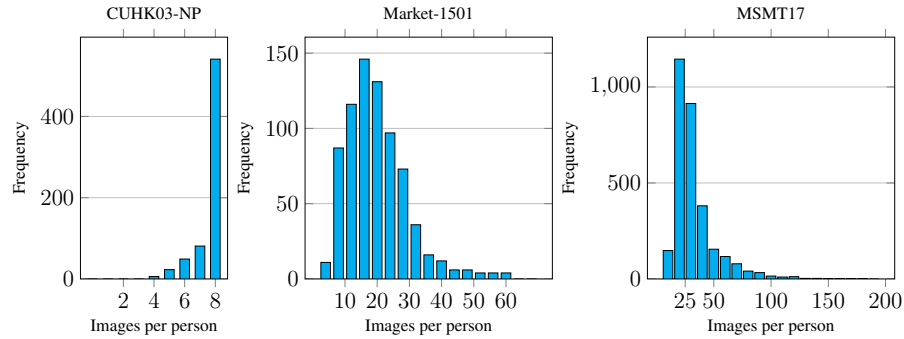


Fig. 6: A histogram of the number of images per person in the gallery set.

| Dataset | CUHK03-NP |      |      |      | Market-1501 |       |       |      | MSMT17 |       |       |      |
|---------|-----------|------|------|------|-------------|-------|-------|------|--------|-------|-------|------|
| Subset  | 4-5       | 6    | 7    | 8    | 1-15        | 16-30 | 31-50 | 50+  | 1-10   | 11-20 | 21-30 | 30+  |
| mAP     | 91.9      | 92.1 | 92.2 | 92.3 | 80.2        | 84.4  | 83.3  | 80.5 | 96.1   | 97.5  | 97.2  | 97.1 |

Table 3: The mAP for different subsets of the CUHK03-NP, Market-1501, and MSMT17 datasets.

quires a larger number of images to construct a comprehensive representation of an individual compared to the Market-1501 dataset.

**Qualitative evaluation.** In Fig. 7, we display the top four retrieved images for queries using different methods. For GCP, the retrieved images correspond to those closest to the prototypes in the feature space. Both the people in white and yellow shirts are successfully identified at R-1 by the GCP model (third row). The fourth row illustrates the limitations of the baseline  $\alpha$ -FPS algorithm, e.g., where it fails to retrieve the person with the correct identity. Nevertheless,  $\alpha$ -FPS performs better than the instance-based and centroid-based methods in the case of identifying the person in the white shirt.

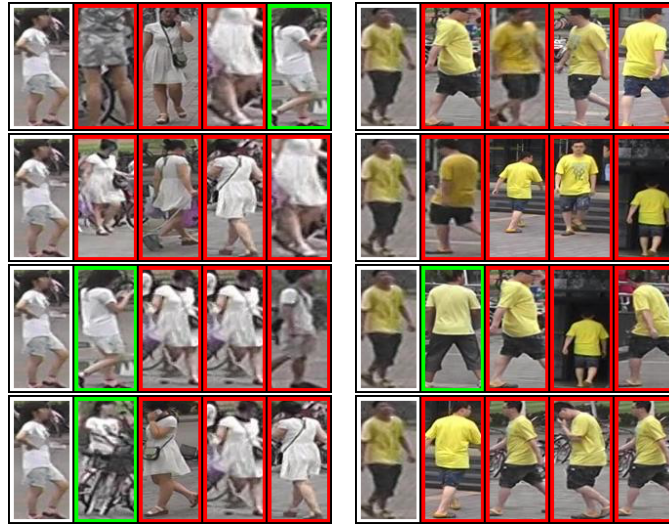


Fig. 7: Examples of retrieved person Re-ID images. The first two rows show the results of instance-based and centroid-based methods, respectively. The images in the third and fourth rows are the results of our attention-based GCP and  $\alpha$ -FPS methods, respectively. In each row, the query images are enclosed by white borders. Images to the right of the query display the top four retrieved images. The green/red borders indicate whether the images share the same/different identities as the query.

To qualitatively examine the use of GCPs, we selected all gallery images of an arbitrary class to generate  $N$  prototypes in sequence. Since each prototype represents only a point in feature space, we tag each with its nearest gallery image in the feature space to provide a visual reference. The generated prototypes and their corresponding

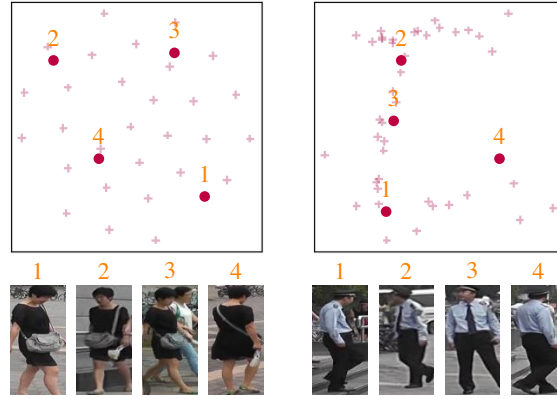


Fig. 8: Prototypes generated using GCPs with their nearest gallery images in the feature space. The number above each prototype indicates the  $i^{th}$  iteration in which it was generated.

tagged images are shown in Fig. 8. The GCP approach is able to effectively identify key images that cover the feature space.

#### 4.4 Discussion

The number of prototypes is a critical parameter in the GCP model. We conducted several experiments that vary the number of prototypes per class. As shown in Table 4, increasing the number of prototypes per class enhances accuracy, but decreases mAP. This is because a higher number of prototypes shifts the approach towards instance-based Re-ID. Unless stated otherwise, all results reported in this paper use  $N = 3$ .

| $N$ | CUHK03-NP   |             | Market-1501 |             |
|-----|-------------|-------------|-------------|-------------|
|     | R-1         | mAP         | R-1         | mAP         |
| 2   | 92.2        | 92.2        | 97.1        | <b>97.2</b> |
| 3   | 93.1        | <b>92.2</b> | 97.3        | 97.1        |
| 4   | 93.5        | 91.9        | 97.3        | 97.0        |
| 5   | 94.0        | 91.7        | <b>97.5</b> | 97.0        |
| 6   | <b>94.3</b> | 91.4        | 97.5        | 96.8        |

Table 4: The effect of the number of prototypes ( $N$ ) per class on the performance using the CUHK03-NP and Market-1501 datasets.

It is worth noting that during the inference stage, the class information of the gallery images is utilized to group them and generate prototypes. Exposure to such information is common in practical applications. For example, in surveillance scenarios, multiple images of a target individual are often known in advance and readily available for comparison against the query image.

## 5 Conclusion

This paper introduced the concept of a GCP, which encapsulates both instance-based and centroid-based image retrieval. To do this, we first analyzed the intricacies of various types of class representatives. Then, we introduced a learning-based architecture to generate robust class prototypes. Additionally, we developed a straightforward yet effective algorithm,  $\alpha$ -FPS, as a baseline method for selecting prototypes without requiring model training. Experimental results show that our GCP approach surpasses modern techniques on person Re-ID benchmark datasets. In future work, we aim to further refine this methodology to identify superior class prototypes within the generalized framework.

## Acknowledgments

This material is based upon work supported by the Air Force Research Laboratory under award number FA8571-23-C-0041.

## References

1. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8351–8361 (2019)
2. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: A deep quadruplet network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 403–412 (2017)
3. Chen, W., Xu, X., Jia, J., Luo, H., Wang, Y., Wang, F., Jin, R., Sun, X.: Beyond appearance: A semantic controllable self-supervised learning framework for human-centric visual tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15050–15061 (2023)
4. Chung, D., Tahboub, K., Delp, E.J.: A two stream siamese convolutional neural network for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1983–1991 (2017)
5. Eldar, Y., Lindenbaum, M., Porat, M., Zeevi, Y.Y.: The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing* **6**(9), 1305–1315 (1997)
6. Ess, A., Leibe, B., Van Gool, L.: Depth and appearance for mobile scene analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–8 (2007)
7. <https://github.com/robotic-vision-lab/Person-Re-Identification-Via-Generalized-Class-Prototypes>
8. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance. vol. 3, pp. 1–7 (2007)
9. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15013–15022 (2021)
10. He, W., Deng, Y., Tang, S., Chen, Q., Xie, Q., Wang, Y., Bai, L., Zhu, F., Zhao, R., Ouyang, W., Qi, D., Yan, Y.: Instruct-reid: A multi-purpose person re-identification task with instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17521–17531 (2024)

11. Kamousi, P., Lazard, S., Maheshwari, A., Wuhler, S.: Analysis of farthest point sampling for approximating geodesics in a graph. *Computational Geometry* **57**, 1–7 (2016)
12. Lagunes-Fortiz, M., Damen, D., Mayol-Cuevas, W.: Centroids triplet network and temporally-consistent embeddings for in-situ object recognition. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. pp. 10796–10802 (2020)
13. Li, H., Wu, G., Zheng, W.S.: Combined depth space based architecture search for person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6729–6738 (2021)
14. Li, S., Bak, S., Carr, P., Wang, X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 369–378 (2018)
15. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 152–159 (2014)
16. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: Occluded person re-identification with part-aware transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2898–2907 (2021)
17. Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing* **26**(7), 3492–3506 (2017)
18. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5790–5799 (2017)
19. 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 Transactions on Multimedia* **22**(10), 2597–2609 (2019)
20. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. *Journal of Machine Learning Research* **9**(11) (2008)
21. Quan, R., Dong, X., Wu, Y., Zhu, L., Yang, Y.: Auto-reid: Searching for a part-aware convnet for person re-identification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3750–3759 (2019)
22. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5363–5372 (2018)
23. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: *Proceedings of the Advances in Neural Information Processing Systems*. vol. 30 (2017)
24. Somers, V., De Vleeschouwer, C., Alahi, A.: Body part-based representation learning for occluded person re-identification. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 1613–1623 (2023)
25. Tan, H., Liu, X., Bian, Y., Wang, H., Yin, B.: Incomplete descriptor mining with elastic loss for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(1), 160–171 (2021)
26. Vapnik, V.: *The nature of statistical learning theory*. Springer science & business media (2013)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: *Proceedings of the Advances in Neural Information Processing Systems*. vol. 30 (2017)
28. Wang, G., Lai, J., Huang, P., Xie, X.: Spatial-temporal person re-identification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 33, pp. 8933–8940 (2019)

29. Wang, H., Shen, J., Liu, Y., Gao, Y., Gavves, E.: Nformer: Robust person re-identification with neighbor transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7297–7307 (2022)
30. Wang, J., Wang, K.C., Law, M.T., Rudzicz, F., Brudno, M.: Centroid-based deep metric learning for speaker recognition. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 3652–3656 (2019)
31. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 79–88 (2018)
32. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: *Proceedings of the European Conference on Computer Vision*. pp. 499–515. Springer (2016)
33. Wiczarek, M., Rychalska, B., Dabrowski, J.: On the unreasonable effectiveness of centroids in image retrieval. In: *Proceedings of the International Conference on Neural Information Processing*. pp. 212–223. Springer (2021)
34. Yuan, Y., Chen, W., Yang, Y., Wang, Z.: In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 354–355 (2020)
35. Zhang, A., Gao, Y., Niu, Y., Liu, W., Zhou, Y.: Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 598–607 (2021)
36. Zhang, G., Zhang, Y., Zhang, T., Li, B., Pu, S.: Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 14133–14142 (2023)
37. Zhang, Z., Lan, C., Zeng, W., Chen, Z., Chang, S.F.: Beyond triplet loss: Meta prototypical n-tuple loss for person re-identification. *IEEE Transactions on Multimedia* **24**, 4158–4169 (2021)
38. Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal person re-identification via multi-loss dynamic training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8514–8522 (2019)
39. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1116–1124 (2015)
40. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1318–1327 (2017)
41. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5157–5166 (2018)
42. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3702–3712 (2019)
43. Zhou, X., Zhong, Y., Cheng, Z., Liang, F., Ma, L.: Adaptive sparse pairwise loss for object re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19691–19701 (2023)
44. Zhu, Y., Yang, Z., Wang, L., Zhao, S., Hu, X., Tao, D.: Hetero-center loss for cross-modality person re-identification. *Neurocomputing* **386**, 97–109 (2020)

## Appendix

In this appendix we provide additional details on the GCP methodology.

### A Rationale Behind the Effectiveness of GCP

To demonstrate how GCP enhances Re-ID performance, we present a toy example. Consider the distribution of feature vectors for two neighboring classes in the embedding space as depicted in Fig. 9. In Fig. 9a, the gray dot denotes the query feature vector, whose true class label is orange. If we compute the precision of this query image using the five nearest neighbors, the precision will be 0.6 (three true positives and two false positives). Conversely, if we employ a centroid prototype to represent the class instead of using all the individual feature vectors, we can achieve a precision of 1.0 (Fig. 9b). Thus, using the centroid prototype can improve the precision for certain distributions of gallery features.

However, centroid prototypes are not robust to arbitrary image distributions. Consider a different class distribution as illustrated in Fig. 9c. Unlike the scenario in Fig. 9b, many gallery images of the blue class are closer to the orange class. Hence, the blue class centroid is shifted towards the orange class. For such a distribution, if we use the centroid prototype as the class representative and calculate precision based on it, the precision will be very low and the accuracy will be reduced as well. This example underscores the limitation of using a centroid vector as a class representative, it is not robust to arbitrary image distributions. Our GCP method addresses this impediment by considering the gallery image distribution, Fig. 9d.

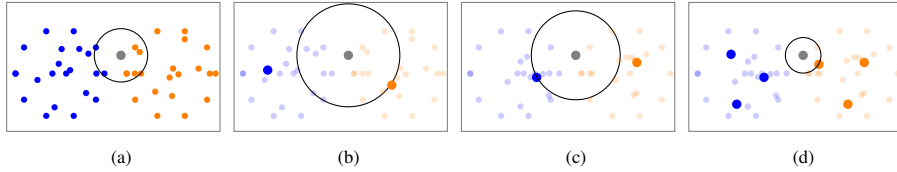


Fig. 9: A comparison between (a) instance-based, (b) centroid-based, and (d) prototype-based Re-ID.

### B $\alpha$ -Farthest Point Sampling

The input to Algorithm 1 is the set  $X$  of class feature vectors, the number  $N$  of prototypes to select, and the parameter  $\alpha$ . By tuning  $\alpha$  we can regulate the degree of interpolation between the selected farthest point and the prior prototype. The output of the algorithm is a set of prototypes  $P$  representing the given class. Concretely, the algorithm starts with a single prototype, which is the centroid of the given feature vectors (line 1). Next, it iteratively selects a point  $x \in X$  that is farthest from the currently selected set of prototypes  $P$ , followed by a  $p \in P$  that is closest to  $x$  (line 3). Subsequently,  $x$  is removed from the available set of features  $X$  (line 4). The algorithm then modifies  $x$  based on the given  $\alpha$  value (line 5) and adds it to the set of prototypes  $P$  (line 6).

## C Convergence Analysis

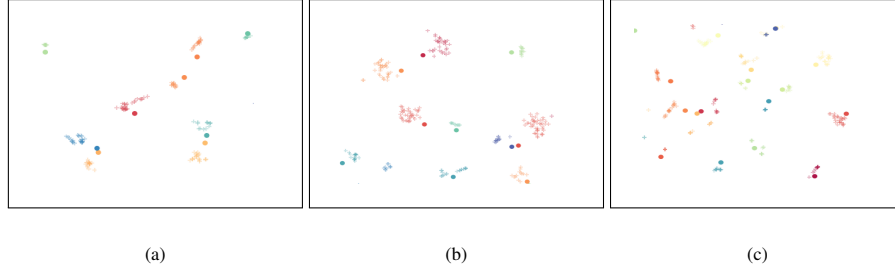


Fig. 10: A subset of the feature vectors (+) in the embedding space for the (a) CUHK03-NP, (b) Market-1501, and (c) MSMT17 datasets. The colored dots represent the class prototypes selected by the GCP method. Best viewed zoomed in.

We sought to investigate whether GCP converges to a centroid-based approach when  $N = 1$ , i.e., when only a single prototype is generated per class. Specifically, we aimed to determine whether the prototype produced by GCP closely resembles the centroid of the class in terms of proximity. To explore this, we conducted multiple experiments across different datasets. Interestingly, GCP did not select the centroid as the prototype in most cases. Instead, it generated prototypes influenced by the distribution of images in the embedding space, which were often located far from the class centroids. To illustrate this, we visualized the feature vectors in a 2D space using t-SNE [20] as shown in Fig. 10. Additionally, we projected the prototypes onto the same embedding space, highlighting their significant distances from the class centroids.

## D Failure Cases

Fig. 11 presents examples where GCP fails to retrieve the correct gallery instance corresponding to the given query. Despite these failures, noticeable visual similarities can be observed between the query images and the retrieved results. In some instances (e.g., the middle column of Fig. 11a and Fig. 11b), distinguishing between the query and retrieved persons is challenging, even for a human.



Fig. 11: Examples where GCP fails at R-1 retrieval on the (a) CUHK02-NP, (b) Market1501, and (c) MSMT17 datasets. The top row displays the query images, while the images directly below represent the R-1 retrievals from the gallery.