# Text-based Person Search via Multi-Granularity Embedding Learning

**Chengji Wang**[1] , **Zhiming Luo**[1*] , **Yaojin Lin**[2] and **Shaozi Li**[1*]
[1]Department of Artificial Intelligence, Xiamen University, China
[2]School of Computer Science, Minnan Normal University, China
chenjw@stu.xmu.edu.cn, {zhiming.luo,szlig}@xmu.edu.cn, yjlin@mnnu.edu.cn

## Abstract

Most existing text-based person search methods highly depend on exploring the corresponding relations between the regions of the image and the words in the sentence. However, these methods correlated image regions and words in the same semantic granularity. It 1) results in irrelevant corresponding relations between image and text, 2) causes an ambiguity embedding problem. In this study, we propose a novel multi-granularity embedding learning model for text-based person search. It generates multi-granularity embeddings of partial person bodies in a coarse-to-fine manner by revisiting the person image at different spatial scales. Specifically, we distill the partial knowledge from image scrips to guide the model to select the semantically relevant words from the text description. It can learn discriminative and modality-invariant visual-textual embeddings. In addition, we integrate the partial embeddings at each granularity and perform multi-granularity image-text matching. Extensive experiments validate the effectiveness of our method, which can achieve new state-of-the-art performance by the learned discriminative partial embeddings.

## 1 Introduction

Text-based person search [Li *et al.*, 2017b] aims to retrieve the corresponding person images by a textual description from the large-scale image gallery. With the explosive increase of surveillance videos, automatic person searches are urgently demanded for constructing intelligent surveillance systems. Textual descriptions are easily accessible and can describe more details in a more natural way, and text-based person search has gained more and more attention. On the other hand, text-based person search is a challenging problem. Compared with the image-based person search [Wang *et al.*, 2018b; Rao *et al.*, 2020; Wang *et al.*, 2020], text-based person search further needs to overcome the modality gap between image and text. Different from the conventional image-text matching task [Lee *et al.*, 2018], text-based person search is a fine-grained retrieval problem, that all the images belong to the same category, *i.e.*, pedestrian.

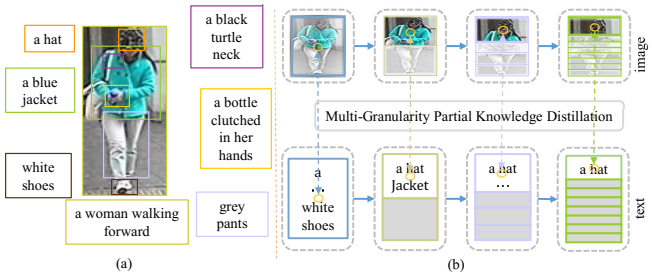



Figure 1: (a) The image regions and words are correlated in different semantic granularities. (b) An example of multi-granularity embeddings. We show a person in a coarse-to-fine manner to learn multi-granularity person representations. In each granularity, the model focuses on the person parts with different spatial scales. We design a multi-granularity partial knowledge distillation loss to learn discrimination and modality-invariant partial representations.

Current state-of-the-art methods [Jing *et al.*, 2020; Zhe *et al.*, 2020; Zheng *et al.*, 2020a] mainly focus on excavating fine-grained local parts and then perform fine-grained visual-textual matching. These methods can learn discriminative partial representations and achieve good performance, but they are troubled by the ambiguity embedding problem. First, image regions and words are correlated in different semantic granularities. The discriminative parts of a person usually occur at different granularities, which is ignored in previous methods and results in an ambiguity embedding problem. As is shown in Figure 1(a), "a jacket with a black turtleneck" and "a blue jacket" describe the upper body of the person in different granularities. The neck and the jacket occupy different size image areas, which should be treated differently. Second, the works before 2020 [Li *et al.*, 2017b; Li *et al.*, 2017a; Chen *et al.*, 2018a] usually utilize the words as guidance to integrate image regions. However, not all words correspond to the image contents. The irrelevant words can mislead the model and aggravate the ambiguity embedding problem. In addition, recent methods [Jing *et al.*, 2020; Zhe *et al.*, 2020] require extra annotations for extracting meaningful image regions that are not always available in a real scenario. Therefore, we are encouraged to explore multi-granularity image-text representations and align the image

*corresponding author

and text in multi-granularity without extra annotations.

To address the ambiguity embedding problem, in this paper, we specifically propose a novel multi-granularity embedding learning (MGEL) model. As is shown in Figure 1, we represent a person in a coarse-to-fine manner and perform multi-granularity knowledge distillation. **First**, we slice the person images into scrips at different spatial scales to extract multi-granularity partial embeddings, in which each member captures the discriminative part representation from different granularities. A pedestrian is from head to foot, each scrip can focus on a specific partial person body. Given a textual description, we adopt a multi-head self-attention module to extract partial embeddings at different granularities. **Second**, we design a part alignment loss to distill the partial knowledge of a person from image to guide learned discriminative multi-granularity textual partial embeddings. This loss can 1) reduce the modality-gap between image and text, 2) guide the self-attention module to automatically filter out irrelevant words which is helpful to solve the ambiguity partial embedding problem. **Third**, we aggregate the partial embeddings to obtain a global person embedding, our proposed MGEL model performs multi-granularity image-text matching.

In summary, the contributions of this work are three-fold: **First**, we propose a multi-granularity embedding learning (MGEL) model to exploit and integrate multi-granularity partial knowledge for text-based person search. It enables learning more discriminative and modality-invariant part features. **Second**, we develop a multi-granularity knowledge distillation-based part alignment loss to solve the ambiguity embedding problem. **Finally**, we conduct extensive and fair experiments on a public benchmark to demonstrate the effectiveness of the proposed method. Our proposed MGEL model significantly outperforms previous methods with the same visual backbone, and achieves a new state of the art.

## 2 Related Works

### 2.1 Text-based Person Search

Text-based person search is introduced by Li et al [Li *et al.*, 2017b], they collect a large-scale person description dataset, *CUHK-PEDES* and design a Recurrent Neural Network with Gated Neural Attention mechanism model (GNA-RNN) for this task. Most of the following works [Li *et al.*, 2017a; Chen *et al.*, 2018b; Chen *et al.*, 2018a; Jing *et al.*, 2020; Zheng *et al.*, 2020a] adopt the cross-modality attention mechanism to attend all the image regions of images and the corresponding words in textual description, the core idea is to obtain weighted matching between image and text for alleviating the irrelevant matching. These methods are inefficient and increase the complexity of computation. The joint embedding based methods [Zhang and Lu, 2018; Sarafianos *et al.*, 2019; Zheng *et al.*, 2020b] directly compute the matching score for image-text pairs in a shared latent space. These methods are computation efficient at the test stage, but ignore the part representations which play a key role in text-based person search. There are also some attempts at adversarial learning to reduce modality-gap [Liu *et al.*, 2019; Sarafianos *et al.*, 2019]. Recently, more and more works focus on extract fine-grained representations, *e.g.*, person attributes [Aggarwal *et al.*, 2020; Zhe *et al.*, 2020], the representations of human keypoints [Jing *et al.*, 2020]. However, the above methods require extra annotations.

### 2.2 Part-based Re-ID Methods

Recently, many deep person re-identification methods improve the performance by exploring fine-grained and discriminative part features of person [Sun *et al.*, 2018; Fu *et al.*, 2018; Wang *et al.*, 2018a; Zheng *et al.*, 2019]. These methods mainly divide the person images into multiple spatial bins, compute the part-level representations, and utilize extra loss functions for training each part. Although these methods have demonstrated their effectiveness in the person re-ID task, the same technique can not directly be applied to the text-based person search task since we cannot divide the text description as the same as the image.

In this study, we have modified the above methods to apply it to text-based person search. We deal with the images and texts separately. We slice the images into spatial bins at different spatial scales. For the text, we adopt multi-head self-attention to select semantic relevant words. A multi-granularity knowledge distillation based part alignment loss is proposed and used.

## 3 Approach

In this section, we explain our proposed multi-granularity embedding learning (MGEL) method in detail. Our objective is to learn multi-granularity and discriminative embeddings of partial person bodies, and the overall architecture is shown in Figure 2. We first fed the image into the CNN backbone network to extract the image feature maps and also utilize a Bi-LSTM to obtain the word embeddings. Then, we slice the image feature maps in a horizontal manner at various image scales to generate multi-granularity part scrips. For the text stream, we adopt the multi-head self attention module to extract the partial embeddings of parts at different granularities. Because of the specific distribution of pedestrian images, each image scrip roughly corresponds to a special person part. We design a multi-granularity knowledge distillation-based part alignment loss to adaptively select semantic relevant words from text. Finally, we perform multi-granularity image-text matching at the global level by aggregating the partial embeddings at each granularity. Our core contribution is multi-granularity embedding learning. It 1) represents the image and text in a coarse-to-fine manner, 2) performs multi-granularity visual-textual alignment, 3) distills the knowledge of parts from image to guide learning corresponding partial embeddings from text.

### 3.1 Multi-Granularity Partial Embedding Learning

#### Modality-Aware Partial Embedding

Denoting the feature maps of image $x$ extracted by the backbone network as $X \in \mathbb{R}^{c \times H \times W}$. We extract features at 4 granularities in the MGEL model and $X$ is sliced into several spatial bins horizontally at each granularity. Specifically, we design a granularity attention block (GAB), as is shown in Figure 2, to eliminate the distraction of unrelated

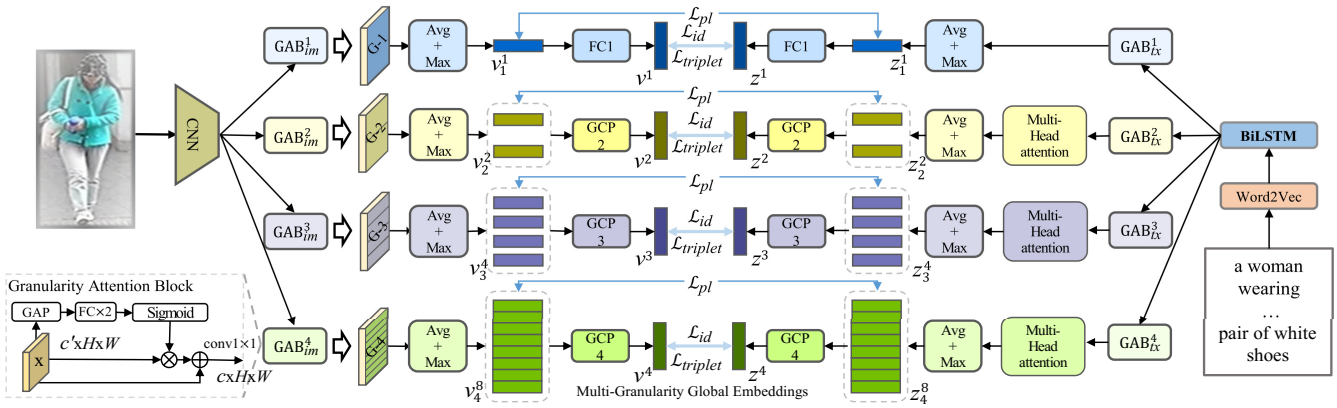



Figure 2: The overall architecture of the proposed multi-granularity embedding learning model. It consists of 4 granularities (G-1, G-2, G-3, G-4), and performs multi-granularity multi-level embedding learning and multi-granularity cross-modality matching. The GAB is the granularity attention block and GCP is global contrastive pooling. G is the abbreviation of granularity.

features, it contains a channel attention block and $1 \times 1$ convolutional layer. For the $k^{th}$ granularity, we have $\mathbf{X}^k = conv(\mathrm{ca}(X) + X)$, where $\mathrm{ca}()$ indicates the channel attention. After that, we slice the feature maps into $2^{k-1}$ bins with each bin $\mathbf{X}_i^k \in \mathbb{R}^{c \times \frac{H}{2^{k-1}} \times W}$. $i$ stands for the index of bins in each granularity. For instance, $\mathbf{X}_1^4$ means the first bin in the fourth granularity. Then, we pool each spatial bin $\mathbf{X}_i^k$ by global average pooling and max pooling to generate the vector representation, $v_i^k$

$$v_i^k = avgpool(\mathbf{X}_i^k) + maxpool(\mathbf{X}_i^k). \tag{1}$$

The $\mathbf{v}_i^k$ is the $i^{th}$ partial embedding at $k^{th}$ different granularity. The multi-granularity partial embeddings of image $x$ can be represented as $\{v_i^k, k \in \{1,2,3,4\}, i \in \{1, \cdots, 2^{k-1}\}\} \in \mathbb{R}^c$.

Given the word embeddings $Z \in \mathbb{R}^{c \times T}$ of a text $z$ with $T$ words, the granularity attention block is firstly applied to extract granularity related features $\mathbf{Z}^k \in \mathbb{R}^{c \times T}$. Then, for the $k^{th}$ $(k > 1)$ granularity, the multi-head self-attention module takes the granularity related features as an input and outputs $2^{k-1}$ partial embeddings,

$$\mathbf{Z}_i^k = \left[\mathrm{softmax}(P_i^k \mathbf{Z}^k)^T (Q_i^k \mathbf{Z}^k)\right] (U_i^k \mathbf{Z}^k)^T \tag{2}$$

where $P_i^k \in \mathbb{R}^{c \times c}$, $Q_i^k \in \mathbb{R}^{c \times c}$, $U_i^k \in \mathbb{R}^{c \times c}$. After that, we use the pooling methods in Eq. 1 to obtain the vector representation $z_i^k$ of $i^{th}$ person part in $k^{th}$ granularity. For the first granularity, we directly perform a global pooling to obtain the partial embeddings $z_1^1$. The final multi-granularity textual embeddings can be represented as $\left\{z_i^k, k \in \{1,2,3,4\}, i \in \{1, \cdots, 2^{k-1}\}\right\} \in \mathbb{R}^c$.

**Partial Embedding Learning**

In this section, we aim to correspond the multi-granularity partial embeddings of text to the image scrips for learning discriminative partial embeddings. We accomplish this by considering the following two aspects: 1) the visual embedding and textual embedding of the same partial body should have the same statistical distribution; 2) the two embeddings need to be modality-invariant.

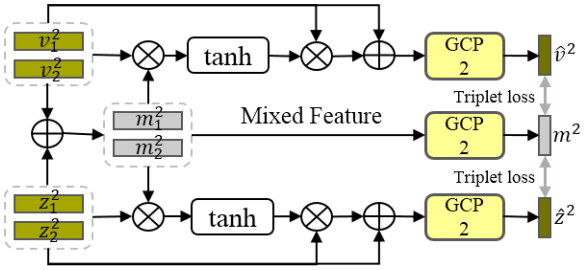



Figure 3: An visual example of the mixed feature alignment.

Given the image partial embeddings $\{v_i^k\}$ and text partial embeddings $\{z_i^k\}$ at the $k^{th}$ granularity, we first adopt an activation mapping function [Zagoruyko and Komodakis, 2017] across the channel dimension to obtain the activate values,

$$\hat{A}_i^k = \frac{1}{c}\sum |v_i^k| \in \mathbb{R} \text{ and } \check{A}_i^k = \frac{1}{c}\sum |z_i^k| \in \mathbb{R}, \tag{3}$$

where $|\cdot|$ indicates the abs function. Then, we minimize

$$\mathcal{L}_{pull}^k = \sum_{i=1}^{2^{k-1}} \|\hat{A}_i^k - \check{A}_i^k\|^2. \tag{4}$$

For the other aspect, the $v_i^k$ and $z_i^k$ correspond to the same parts, their mixed feature also represents the same part. As shown in Figure 3, we fuse the two modality features by an average pooling:

$$m_i^k = \frac{1}{2}(v_i^k + z_i^k). \tag{5}$$

Then, we perform element-wise product between three features to model the co-interaction between two modality and utilize tanh function to generate modality-specified co-attentive features. In short, the attentive partial embeddings can be represented by

$$\begin{aligned} \hat{v}_i^k &= v_i^k \cdot \frac{1}{2}(1 + \tanh(v_i^k \cdot m_i^k)), \\ \hat{z}_i^k &= z_i^k \cdot \frac{1}{2}(1 + \tanh(z_i^k \cdot m_i^k)). \end{aligned} \tag{6}$$

The element-wise product between $v_i^k$ (or $z_j^k$) and $m_i^k$ can leverage $m_i^k$ as guidance to find where the feature from each modality is attended by the communal feature. The element-wise product can estimate the co-interaction between features. Optimizing features generated by element-wise products can reduce inter-modality discrepancy.

We aggregate the partial embeddings by global contrastive pooling (GCP) [Park and Ham, 2020]. For example, given the attentive partial image embeddings $\hat{v}_i^k$, we firstly perform average and max pooling, resulting in features $\hat{v}_a^k$ and $\hat{v}_m^k$. Then we 1) fed the $\hat{v}_m^k$ into a fully connected layer and get feature $\overline{v}_m^k$, 2) compute a contrastive feature by subtracting $\hat{v}_m^k$ from $\hat{v}_a^k$ and fed it into a fully connected layer. We concatenate $\overline{v}_m^k$ and contrastive feature to be a long vector, then transform it to be $c$ dimension. Finally, we add this vector on $\overline{v}_m^k$ to get final global embedding

$$\hat{v}^k = fc_1(\hat{v}_m^k) + fc_3([fc_1(\hat{v}_m^k), fc2(\hat{v}_m^k - \hat{v}_a^k)]), \quad (7)$$

where $[\cdot,\cdot]$ is the concatenation operation, $fc_1$, $fc_2$ and $fc_3$ are fully connected layers with $c$ dimensional outputs. Similarly, given the attentive partial text embedding $\hat{z}_i^k$ and mixed embedding $m_i^k$, we can get the global embeddings $\hat{z}^k$ and $\hat{m}^k$. The parameters of GCP are shareable at each granularity. The overall process is illustrated in Figure 3. For encouraging the two embedding to be modality-invariant, we use the triplet loss for this step,

$$\begin{aligned}\mathcal{L}_{mix}^k = &\max[0, \alpha - \cos(\hat{x}^k, m^k) + \cos(\hat{x}^k, \hat{m}^k)], \\ &+ \max[0, \alpha - \cos(\hat{z}^k, m^k) + \cos(\hat{z}^k, \check{m}^k)],\end{aligned} \quad (8)$$

where $\hat{m}^k$, $\check{m}^k$ are the hardest $m^k$ for $\hat{x}^k$ and $\hat{z}^k$, respectively.

By jointly considering these two aspects, the final part alignment loss for this partial embedding learning is

$$\mathcal{L}_{pl}^k = \mathcal{L}_{pull}^k + \mathcal{L}_{mix}^k. \quad (9)$$

### 3.2 Multi-Granularity Image-Text Matching

**Multi-Granularity Global Embedding**

Although the partial embeddings contain rich part information, they are still limited for discriminatively representing the identity of a person. We aggregate the part representations at each granularity by the global contrastive pooling (GCP) [Park and Ham, 2020] and generate multi-granularity global person embeddings. Given the text embeddings $\{z_i^k\}$ from $k^{th}$ granularity, we have the global text embedding $z^k$. Given the image embeddings $\{v_i^k\}$ from $k^{th}$ granularity, we have the global image embedding $v^k$. For the first granularity, we directly apply a cross-modality parameter shared fully connected layer to obtain the global embeddings. After the above processing, we have the global image embeddings $\{v^k, k \in \{1,2,3,4\}\}$ and global text embeddings $\{z^k, k \in \{1,2,3,4\}\}$.

**Image-Text Matching**

The triplet loss is a common objective function for the retrieval task. In this study, we employ the triplet loss to train our MGEL model. For the $k^{th}$ granularity, we have

$$\begin{aligned}\mathcal{L}_{triplet}^k = &\max[0, \alpha - \cos(v^k, z^k) + \cos(v^k, \tilde{z}^k)], \\ &+ \max[0, \alpha - \cos(z^k, v^k) + \cos(z^k, \tilde{v}^k)],\end{aligned} \quad (10)$$

where $\tilde{z}^k$ and $\tilde{v}^k$ are the hardest negative samples within a training batch.

We also adopt the cross-entropy loss to match the person in identity-level, ensuring the image and text with the same identity have similar representations. We first get the predicted identity probabilities $\hat{y}^k$ of image $x$ and $\check{y}^k$ of text $z$ by a shared softmax layer. Then, we have the cross-entropy loss,

$$\mathcal{L}_{id}^k = -y_{id}\log(\hat{y}^k) - y_{id}\log(\check{y}^k). \quad (11)$$

where $y_{id}$ is the one-hot identity label of $x$ and $z$.

### 3.3 Training and Testing

During the training phase, we optimize the proposed model by jointly considering the triplet loss, identity classification loss and partial embedding learning loss. The final loss is denoted as,

$$\mathcal{L} = \sum_{k=1}^{4} \mathcal{L}_{triplet}^k + \mathcal{L}_{id}^k + \mathcal{L}_{pl}^k. \quad (12)$$

In the testing stage, we concatenate the $4$ global embeddings into a long vector for each modality, and leverage the cosine distance to measure the similarity between image and text samples.

## 4 Experiments

### 4.1 Experiment Setup

**Datasets.** *CUHK-PEDES* dataset [Li *et al.*, 2017b] is a large-scale text-based person search dataset that contains 40,206 images from 13,003 identities. Each image is roughly associated with two different text descriptions. As same as [Li *et al.*, 2017b; Zhang and Lu, 2018], we split the dataset into three subsets: training set, validation set, and testing set. The person identities of these three subsets are disjoint. The training set includes 34,054 images and 68,126 textual descriptions of 11,003 persons. The validation set has 3,078 images and 6,158 textual descriptions of 1,000 persons. The testing set contains 3,074 images and 6,156 textual descriptions of 1,000 persons.

**Evaluation metrics.** We adopt the widely used Rank-k (k=1, 5, 10) and mean Average Precision (mAP) to evaluate the performance. Given the query textual descriptions, the Rank-k indicates the percentage of the queries with at least one correct item among their top-k retrieved results. The mAP is the mean of the average precision scores for each query.

**Implementation details.** The vocabulary includes 12,000 words, and we represent each word by a 300-dimension vector. The feature dimension $c$ is 512. We use Adam optimization to train the model with a learning rate of $2e-4$. All the models are trained with 50 epochs and a mini-batch contains 32 image-text pairs.

### 4.2 Comparison with State-of-the-Art Methods

In Table 1, we report the performance comparison of the proposed MGEL against the state-of-the-art methods in terms of Rank-k accuracy in the text-to-image retrieval task. The

| Visual | Method | Venue & Year | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|---|
| VGG-16 | GNA-RNN | CVPR 2016 | 19.05 | - | 53.64 |
| | CMCE | ICCV 2017 | 25.94 | - | 60.48 |
| | PWM-ATH | WACV 2018 | 27.14 | 49.45 | 61.02 |
| | GLA | ECCV 2018 | 43.58 | 66.93 | 76.26 |
| | Dual-Path | TOMM 2020 | 32.15 | 54.42 | 64.30 |
| | GLAM | AAAI 2020 | 47.82 | 69.83 | 78.31 |
| | MGEL(ours) | - | **52.68** | **74.37** | **83.11** |
| MobileNet | CMPC+CMPM | ECCV 2018 | 49.37 | 71.69 | 79.27 |
| | CMAAM | WACV 2020 | 55.13 | 76.14 | 83.77 |
| | MGEL(ours) | - | **59.21** | **79.16** | **85.88** |
| ResNet-50 | Dual-Path | TOMM 2020 | 44.40 | 66.26 | 75.07 |
| | MIA | TIP 2020 | 53.10 | 75.00 | 82.90 |
| | A-GANet | ACM MM 2019 | 53.14 | 74.03 | 82.95 |
| | GLAM | AAAI 2020 | 54.12 | 75.45 | 82.97 |
| | ViTAA | ECCV 2020 | 55.97 | 75.84 | 83.52 |
| | MGEL(ours) | - | **60.27** | **80.01** | **86.74** |

Table 1: Comparison with state-of-the-art methods on the *CUHK-PEDES* test set on Text-to-Image task.

| Model | Text-to-Image | | Image-to-Text | |
|---|---|---|---|---|
| | Rank-1 | mAP | Rank-1 | mAP |
| No Multi-Granularity | 54.73 | 34.58 | 67.40 | 33.53 |
| No Global Embedding | 55.29 | 35.25 | 67.93 | 33.79 |
| MGEL | **59.21** | **37.62** | **70.88** | **36.57** |

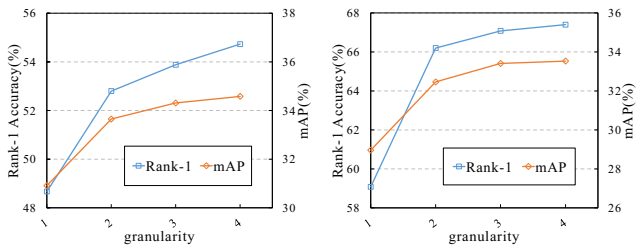
Table 2: Evaluation of the Multi-Granularity and Multi-Level Embedding structure. No Multi-Granularity model uses the fourth granularity (G-4, features $v^4$ and $z^4$). No Global Embedding model has 4 granularities and directly concatenates the partial embeddings to perform cross-modality retrieval.

compared methods include GNA-RNN [Li *et al.*, 2017b], CMCE [Li *et al.*, 2017a], PWM-ATH [Chen *et al.*, 2018b], GLA [Chen *et al.*, 2018a], Dual-Path [Zheng *et al.*, 2020b], GLAM [Jing *et al.*, 2020], CMPC+CMPM [Zhang and Lu, 2018], CMAAM [Aggarwal *et al.*, 2020], MIA [Niu *et al.*, 2020], A-GANet [Liu *et al.*, 2019], ViTAA [Zhe *et al.*, 2020].

We evaluate our method with three visual backbones, *i.e.*, VGG [Simonyan and Zisserman, 2015], MobileNet [Howard *et al.*, 2017] and ResNet-50 [He *et al.*, 2016]. Our approach achieves better performance on the *CUHK-PEDES* datasets compared with current leading methods when using the same visual backbone. For example, the MGEL achieves 52.68%, 59.21% and 60.27% Rank-1 accuracy in text-to-image retrieval task with three backbones, respectively. Compared to ViTAA that extracts the representations of person parts with a pre-trained human parsing predictor, our MGEL improves the accuracy of all metrics by a large margin in this task. These results demonstrate that the partitioning strategy in our MGEL can capture more discriminative fine-grained visual-textual information. Moreover, MGEL has the merit of not using extra annotations. From Table 1, we can see that MGEL outperforms the methods, including A-GANet, GLAM, CMAAM and ViTAA, utilizing extra annotations by a large margin that indicates the effectiveness of the proposed MGEL method.

## 4.3 Ablation Studies

To verify the effectiveness of the components in our MGEL model, we design several ablation studies under different settings on *CUHK-PEDES*, ie., the effectiveness of the multi-granularity and multi-level structure, the contribution of each




(a) Text-to-Image (b) Image-to-Text

Figure 4: Evaluation on the effectiveness of 4 granularities on two retrieval tasks. The Rank-1 accuracy and mAP are compared.

| Model | Feature Dim | Text-to-Image | | Image-to-Text | |
|---|---|---|---|---|---|
| | | Rank-1 | mAP | Rank-1 | mAP |
| G-1 | 512 | 48.67 | 30.91 | 59.08 | 28.96 |
| G-12 | 512x2 | 55.90 | 35.72 | 68.80 | 34.86 |
| G-123 | 512x3 | 57.65 | 36.87 | 69.97 | 35.37 |
| G-1234 | 512x4 | **59.21** | **37.62** | **70.88** | **35.57** |

Table 3: Performance comparison of the proposed method with different granularity. G is the abbreviation of granularity. The Rank-1 accuracy and mAP are compared.

granularity, the effectiveness of the Multi-Granularity Structure and the contribution of each component. To ensure fair comparisons, all experiments are conducted with the MobileNet as the backbone.

**Effectiveness of multi-granularity multi-level structure.** We evaluate the Multi-Granularity and Multi-Level Structure. The results are reported in Table 2. **(1)** When removing the multi-granularity structure, we can observe that the Rank-1 accuracy, mAP drops from 59.21% and 37.62 to 54.73% and 34.58% on text-to-image task, respectively. Similarly, on the image-to-text task, the Rank-1 accuracy and mAP also decrease 3.48% and 3.04%, respectively. The multi-granularity structure can align images and text in different granularities. Its effectiveness demonstrates the importance of corresponding image and text in multi-granularity. **(2)** When removing the global embedding, the model directly utilizes partial embeddings to perform image-text matching. From Table 2, we can observe that the Rank-1 accuracy reduces by 3.92% and 2.95% on two tasks, respectively. Moreover, the mAP reduces by 2.37% and 2.78%, respectively. Without the global embedding, the MGEL will focus on partial body parts and ignore the global representation of a person. The global features learned at different granularities are complementary to each other.

**The contribution of each granularity.** We report the Rank-1 accuracy and mAP in the two tasks to show the contribution of each granularity in Figure 4. We can observe that: 1) The performance steadily improves along with the increment of the granularity. The model with G4 can improve the Rank-1 accuracy by 6.06% and 8.32% on two tasks, respectively. The mAP also rises from 30.91% to 34.58 on the image retrieval task and from 28.96% to 33.53 on the text retrieval task. These results demonstrate the effectiveness of the partitioning strategy. With a finer granularity, the model can

| Pooling | Components | | | | | G-4 | | | | G-1234 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Text-to-Image | | Image-to-Text | | Text-to-Image | | Image-to-Text | |
| | Triplet | Identity | PAL | GAB | GCP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 | mAP |
| Avg | ✓ | | | | | 47.22 | 28.40 | 56.91 | 27.49 | - | - | - | - |
| | | ✓ | | | | 8.95 | 5.72 | 7.48 | 3.30 | 38.32 | 24.55 | 29.25 | 12.37 |
| | ✓ | ✓ | | | | 49.16 | 29.64 | 59.72 | 29.33 | 53.02 | 33.60 | 64.77 | 32.47 |
| | ✓ | ✓ | ✓ | | | 51.15 | 31.80 | 61.81 | 30.23 | 54.01 | 34.70 | 66.75 | 33.74 |
| | ✓ | ✓ | ✓ | | ✓ | 54.04 | 34.26 | 65.29 | 31.91 | 55.26 | 35.41 | 68.41 | 34.28 |
| | ✓ | ✓ | ✓ | ✓ | ✓ | 54.22 | 34.56 | 66.10 | 33.09 | 57.02 | 36.77 | 70.30 | 35.87 |
| Max | ✓ | ✓ | ✓ | ✓ | ✓ | 53.77 | 34.28 | 67.08 | 33.45 | 57.96 | 37.10 | 69.62 | 36.04 |
| Avg+Max | ✓ | ✓ | ✓ | ✓ | ✓ | **54.73** | **34.58** | **67.40** | **33.53** | **59.21** | **37.62** | **70.88** | **36.57** |

Table 4: Evaluation of the effectiveness of the components in MGEL with different settings. We also compare different pooling methods.

exploit more partial patterns by learning more combinations of words. Besides, we further try a finer granularity, such as $k = 5$ with 16 scrips. However, it will bring additional computational costs, and there is no noticeable improvement. Therefore, we adopt four granularities in this study.

**Effectiveness of the multi-granularity structure.** Table 3 shows the results of MGEL with different granularities. From Table 3, we can obtain: 1) The multi-granularity structure is effective in that integrating multi-granularity information can significantly and continuously improve the performance. Compared to G-1, the G-12 improves the Rank-1 accuracy from 48.67% to 55.90% and 59.08% to 68.80% on two tasks, the mAPs are improved from 30.91% to 35.72% and 28.96% to 34.86%. G-1234 model improves the Rank-1 accuracy by 10.54% and 10.89% on two tasks, and improves the mAPs by 6.71% and 6.61%. The multi-granularity structure can not only integrate local and global information, but also the gradual transition process between them is also incorporated, which increases the discriminative ability of features. Comparing Table 3 with Figure 4. The models with multi-granularity outperform the model with single granularity by a large margin. For example, compared to G-2, the G-12 improves the Rank-1 accuracy by 2.02% and improves the mAP by 1.15% on text-to-image retrieval task. We also can see the same phenomenons on other models. The above comparisons can further demonstrate the effectiveness of the multi-granularity embedding.

**Components analysis.** To validate the contribution of each component in our model, we carry out several ablation experiments. We start with a fundamental option that implements with single granularity (G-4). Then, we evaluate each component in the model with multi-granularity (G-1234). The results are shown in Table 4. We observe that **(1)** The combination of triplet loss and identification loss in the two models obtain better performance than using them individually. When only using the triplet ranking loss, we find that the G-1234 model can not converge and fails to learn proper partial representations from the text. **(2)** By sensibly adopting the part alignment loss (PAL), the models can learn more discriminative visual-textual embeddings, and we can observe significant performance gain on all metrics on two tasks. For the G-4 and G-1234, it can align the partial embeddings between two modalities. The image scrips can guide the model to filter unrelated words and learn effective partial embeddings from textual descriptions. **(3)** The GCP module can effectively aggregate partial information to obtain discriminative person embeddings and reduce the feature dimension to speed up retrieval. For the model of G-4, we can observe that the GCP module brings significant performance gains on two tasks. For G-1234, it rises the Rank-1 accuracy from 53.02% to 54.01% and 64.77% to 66.75% on two tasks. **(4)** In a multi-granularity setting, the granularity attention block improves the Rank-1 accuracy by 1.76% and 1.89%, and mAP by 1.36% and 1.59%, respectively. These validate our motivation of using GAB to learn granularity-related features. The GAB also can increase the discriminative ability of features, in which we can observe performance gain on G-4. In Table 4, we compare MGEL with different pooling strategies. As can be seen, the average pooling achieves the best performance on the models with a single granularity setting. However, the Avg+Max achieves the best performance on the model with multi-granularity.

## 5 Conclusion

In this paper, we propose a multi-granularity embedding learning (MGEL) model for addressing text-based person search task. The proposed MGEL exploits partial information of each person, which can successfully enhance the discriminative ability of partial features. Besides, integrating these partial features can generate a more robust feature representation for the target person. To be summarized, the advantages of our method are in three-folds. 1) We construct a coarse-to-fine embedding model that exploits the representations of partial person bodies in different granularity. The model generates discriminative embeddings of partial person parts with introducing very few computation costs. 2) The part alignment loss can exploit multi-level semantic relevance between the natural language descriptions and the corresponding visual content from four different granularities. This loss distills knowledge from the image to select semantic relevant words from the text. 3) Our MGEL model relaxes the requirement of detection or parsing models and thus achieves a new state of the art on the benchmark *CUHK-PEDES*.

## Acknowledgements

This work is supported by the National Nature Science Foundation of China (No. 61876159, No. 61806172, No. 62076116, No. U1705286), the China Postdoctoral Science Foundation Grant (No. 2019M652257).

## References


[Aggarwal *et al.*, 2020] Surbhi Aggarwal, R.Venkatesh Babu, and Anirban Chakraborty. Text-based person search via attribute-aided matching. In *Proceedings of WACV*, 2020.

[Chen *et al.*, 2018a] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *Proceedings of ECCV*, 2018.

[Chen *et al.*, 2018b] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *Proceedings of WACV*, 2018.

[Fu *et al.*, 2018] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of AAAI*, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, 2016.

[Howard *et al.*, 2017] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *Proceedings of CVPR*, 2017.

[Jing *et al.*, 2020] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided joint global and attentive local matching network for text-based person search. In *Proceedings of AAAI*, 2020.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of ECCV*, 2018.

[Li *et al.*, 2017a] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of ICCV*, 2017.

[Li *et al.*, 2017b] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of CVPR*, 2017.

[Liu *et al.*, 2019] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. Deep adversarial graph attention convolution network for text-based person search. In *Proceedings of ACM MM*, 2019.

[Niu *et al.*, 2020] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE TIP*, 29, 2020.

[Park and Ham, 2020] Hyunjong Park and Bumsub Ham. Relation network for person re-identification. In *Proceedings of AAAI*, 2020.

[Rao *et al.*, 2020] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Huang Da, Jun Cheng, and Bin Hu. Self-supervised gait encoding with locality-aware attention for person re-identification. In *Proceedings of IJCAI*, 2020.

[Sarafianos *et al.*, 2019] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Adversarial representation learning for text-to-image matching. In *Proceedings of ICCV*, 2019.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR*, 2015.

[Sun *et al.*, 2018] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of ECCV*, 2018.

[Wang *et al.*, 2018a] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of ACM MM*, 2018.

[Wang *et al.*, 2018b] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'Ichi Satoh. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *Proceedings of IJCAI*, 2018.

[Wang *et al.*, 2020] Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, and Shin'ichi Satoh. Beyond intra-modality: A survey of heterogeneous person re-identification. In *Proceedings of IJCAI*, 2020.

[Zagoruyko and Komodakis, 2017] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of ICLR*, 2017.

[Zhang and Lu, 2018] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of ECCV*, 2018.

[Zhe *et al.*, 2020] Wang Zhe, Fang Zhiyuan, Wang Jun, and Yang Yezhou. Vitaa: Visual-textual attributes alignment in person search by natural language. In *Proceedings of ECCV*, 2020.

[Zheng *et al.*, 2019] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of CVPR*, 2019.

[Zheng *et al.*, 2020a] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei. Hierarchical gumbel attention network for text-based person search. In *Proceedings of ACM MM*, 2020.

[Zheng *et al.*, 2020b] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yidong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM TOMM*, 2020.