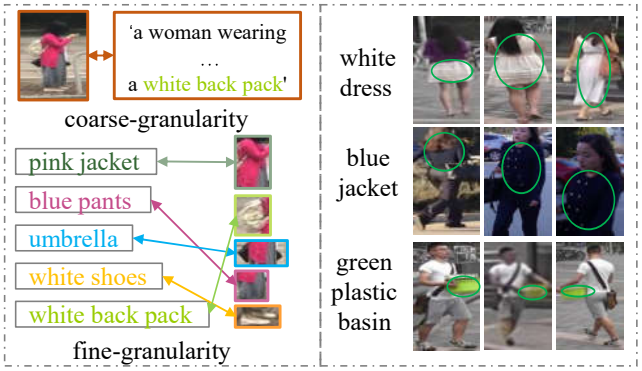
2024 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3503-9015-5/24/$31.00 ©2024 Crown | DOI: 10.1109/ICME57554.2024.10688072

Omni-Granularity Embedding Network for Text-to-Image Person Retrieval

Chengji Wang[1*], Zhiming Luo[2], Shaozi Li[23]
[1]*Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, National Language Resources Monitoring and Research Center for Network Media, School of Computer Science, Central China Normal University, Wuhan, China*
[2]*Department of Artificial Intelligence, Xiamen University, Xiamen, China*
[3]*Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industry, WuyiShan, China*

***Abstract*—Text-to-image person retrieval aims to identify the desired individual based on a textual description. As an instance-level retrieval problem, it has a large intra-class variance and a small inter-class variance. Although significant progress has been made, the omni-granularity matching issue remains unaddressed. Omni-granularity matching involves aligning words with multi-granularity image regions, challenging models to learn in an omni-granularity embedding space. In this paper, we introduce a novel Omni-Granularity Embedding Network (OGEN) for person representation learning. It addresses the omni-granularity matching issue by developing a Cross-Granularity Aggregation Module (CGAM). This module dynamically consolidates diverse granularity features for learning granularity-dependent and omni-granularity person representations. Additionally, a teacher-student knowledge transfer framework is introduced to minimize the inter-modality discrepancy, allowing CGAM to focus on modality-shared semantics. Due to the effectiveness of CGAM and the knowledge transfer framework, our OGEN enhances the Rank-1 accuracy of the Baseline by 8.54%, 9.89%, and 11.09% on three public datasets, respectively.**

***Index Terms*—Text-to-image person retrieval, feature aggregation, knowledge transfer, weight-sharing**



a) Multi-Granularity Matching b) Omni-Granularity Matching

Fig. 1: a) Multi-granularity matching creates numerous embedding spaces for matching images and text at various granularities. b) Omni-granularity matching necessitates aligning words with multi-granularity image regions.

I. Introduction

Text-to-image person retrieval aims to find the interested person from an image gallery based on a natural language description [1]–[3]. Due to the accessibility and flexibility of human-generated natural languages, this task has garnered substantial attention from both academic and industrial domains. Due to the fine-grained retrieval and the modality-gap, text-to-image person retrieval has a large intra-class variance and a small inter-class variance. To tackle these challenges, many effective solutions have been proposed, *e.g.*, global matching [4], [5], local alignment [6], [7], and multi-granularity matching [6], [8]. As is shown in Fig 1(a), multi-granularity matching methods effectively establish fine-grained correspondences between images and text, achieving significant success in text-to-image person retrieval. However, one word may correspond to image patches of dramatically different sizes [8]. As we can see in Fig. 1(b), it is necessary to match "white dress" with multi-granularity image regions. In addition, as the view changes, the style and position of objects are different, *i.e.*,"blue jacket" and "green plastic basin". Multi-granularity matching methods fail to correspond words with multi-granularity image regions.

In this paper, we argue that learned features need to be omni-granularity, which refers to dynamically align words with multi-granularity image regions in a common embedding space. We present Omni-Granularity Embedding Network (OGEN), a novel model designed for learning omni-granularity person representations. By leveraging multi-granularity person embeddings, we develop a divide-and-merge framework, incorporating a novel cross-granularity aggregation module (CGAM) to dynamically combine multi-granularity part representations. The CGAM uses features from one granularity as queries to guide the aggregation of features from another granularity. Its hierarchical structure gradually expands the range of feature interaction, enabling effective information exchange among different granularities. As a result, the generated features exhibit granularity-dependence, thereby achieving omni-granularity.

However, CGAM cannot bridge the modality-gap alone. To overcome the inter-modality difference, we introduce a teacher-student knowledge transfer framework with weight-sharing. The architecture and learnable weights of the teacher

*Corresponding author(wcj@ccnu.edu.cn)

and student networks are shared. The teacher network employs an extra parameter-free dual attention model to enhance cross-modal interaction and bridge the modality-gap. Weight-sharing is an efficient knowledge transfer method between models, allowing the teacher network to impart knowledge to the student through this mechanism. This design enables the student network to concentrate on the shared semantics between the two modalities, thereby improving overall performance and reducing the inter-modality discrepancy.

The main contributions can be summarized as follows: 1) We introduce an omni-granularity embedding network to tackle the omni-granularity matching issue. A novel cross-granularity aggregation module is developed to merge and embed diverse granularity features into an omni-granularity space. 2) We develop a teacher-student knowledge transfer framework to bridge the modality-gap. This framework transfers knowledge based on weight-sharing, guiding the student network to focus on modality-shared semantics. 3) Extensive experiments are conducted on three public benchmark datasets, *i.e.*, CUHK-PEDES [1], ICFG-PEDES [3] and RSTPReid [2] show the effectiveness of our OGEN model.

II. Related Work

Text-to-Image Person Retrieval. Text-to-image person retrieval is introduced by Li et al. [1], who initially collect a large-scale benchmark called CUHK-PEDES dataset. Works [4], [5], [9] jointly embedded image and text into a latent space, aiming to align image and text through global embeddings. Recent works [3], [6]–[8], [10]–[12] focus on constructing fine-grained correspondences between two modalities, attempting to establish word-region correspondences between them. Multi-granularity matching methods [3], [6]–[8], [10], [11] have received the most attention. Due to the varying size of objects in images, multi-granularity matching endeavors to align image and text at different granularities. Image-based person re-identification has addressed omni-granularity matching with works like [13]–[16]. OSNet [13], [16] suggests applying multi-scale convolution to achieve omni-granularity person representation.

Knowledge Transfer. Knowledge transfer refers to a process where a network shares the learned knowledge with others. Weight-sharing is a common method for this. Such as using a pre-trained model for feature extraction in downstream tasks, which has been successful in text-to-image person retrieval [9]. Wen et al. propose sharing the weights of the transformer encoder between image and text, attempting to implicitly achieve cross-modal interaction through weight-sharing [17]. Sung et al. introduce a parameter-efficient Adapter technique by sharing weights to acquire knowledge across tasks, leveraging the sharing of information between tasks on adapters and prompts [18]. Zhao et al. utilize the representation potential of weight-sharing blocks for designing an efficient knowledge distillation mechanism [19].

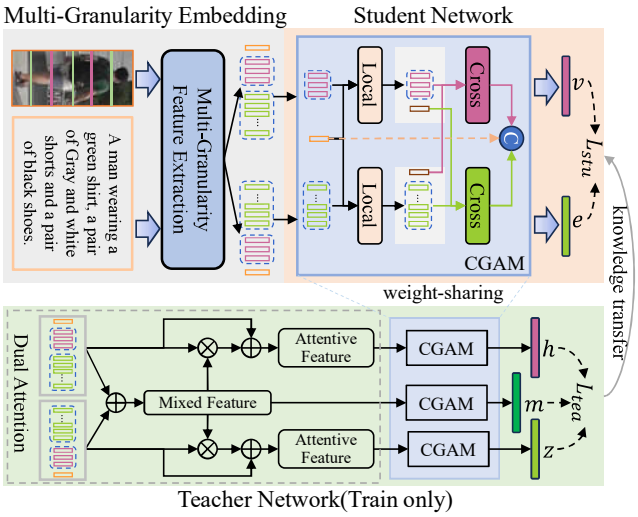


Fig. 2: The framework of our proposed omni-granularity embedding network.

III. Methodology

A. Multi-Granularity Feature Extraction

Visual branch. We employ the popular ViT to extract image features. As depicted in Fig. 2, we divide image into 3 horizontal sections (Granularity 2) and 6 horizontal sections (Granularity 3), respectively. We consolidate the features of corresponding image patches by combining max and average pooling to yield part representations at each granularity. The CLS token output is the granularity 1 representation. We implement an 1×1 convolutional layer and a channel attention module to ensure each granularity attends to distinct features. The resulting multi-granularity person embeddings can be represented as $v_k^s \in R^{1\times d}$, with $k \in \{1, ..., 3\times(s-1)\}$.

Textual Branch. We obtain word embeddings from BERT and a Bi-GRU. We combine the outputs of both directions. Similar to the image branch, we apply a 1×1 convolutional layer and a channel attention module. We use the CLS token as the representation for granularity 1. For granularities $s \in 2, 3$, we employ multi-head attention with $3 \times (s-1)$ heads to extract representations. We acquire multi-granularity embeddings $e_k^s \in R^{1\times d}$ by taking the sum of max and average pooling, where $k \in \{1, ..., 3\times(s-1)\}$.

B. Cross Granularity Aggregation Module

As illustrated in Fig 2, the CGAM module adopts a hierarchical structure. We implement CGAM based on the multi-head attention module (MHA). CGAM considers multi-granularity embeddings as sequences and inputs them into MHA to combine various granularity features. The process of MHA can be represented by:

$$x_{out} = MHA([x_{cls} \| x_{seq}]) \tag{1}$$

where $(\cdot\|\cdot)$ stands for concatenation, the input sequence x_{seq} is concatenated with the CLS token x_{cls}, resulting in a combined representation.

Local Aggregation. The local aggregation treats v^1 from granularity 1 as the CLS token, combining it with other two granularity features to form the input sequence. We then input this sequence into MHA to integrate both global and local information.

$$\begin{aligned} v_{12} &= \mathrm{MHA}(\left[v^1 \| v_k^2\right]) \ (1 \leq k \leq 3), \\ v_{13} &= \mathrm{MHA}(\left[v^1 \| v_k^3\right]) \ (1 \leq k \leq 6), \end{aligned} \tag{2}$$

Then, we enhance the representations using a two-layer feed-forward network and a residual connection.

$$\begin{aligned} v_2 &= \mathrm{FFN}(v_{12}) + v_{12}, \\ v_3 &= \mathrm{FFN}(v_{13}) + v_{13}, \end{aligned} \tag{3}$$

Cross Aggregation. Cross aggregation aims to merge information from three modalities. We swap the first token (CLS) from the local aggregation outputs, and then feed them into two separate MHA modules.

$$\begin{aligned} v^2 &= v_{32} + \mathrm{MHA}(v_{32}), \\ v^3 &= v_{23} + \mathrm{MHA}(v_{23}) \end{aligned} \tag{4}$$

where $v_{32} = \left[v_3[0] \| v_2[1:]\right]$ and $v_{23} = \left[v_2[0] \| v_3[1:]\right]$. After local and cross aggregation blocks, OGAM dynamically integrates multi-granularity features. The omni-granularity image representation is derived by concatenating the CLS tokens of three distinct granularities.

$$v = \left[v^1 \| v^2[0] \| v^3[0]\right] \tag{5}$$

where $v \in \mathbf{R}^{1 \times 3d}$. In the same way, we can obtain the text representations e^1, e^2 and e^3. The omni-granularity text representation is represented by e.

C. Teacher Student Knowledge Transfer

We constructed a weight-sharing teacher-student knowledge transfer framework that utilizes weight-sharing to assist CGAM in suppressing modality-specific information. This enables the teaching student network to learn modality-invariant person representations.

Student Network. As illustrated in Fig. 2, the student network incorporates a cross-modal shared CGAM module. We feed multi-granularity image-text embeddings into the CGAM to extract omni-granularity image-text representations v and e. We employ a bidirectional triplet ranking loss as

$$L_r^o = [\alpha - s(v,e) + s(v,\hat{e})]_+ + [\alpha - s(v,e) + s(\hat{v},e)]_+ \tag{6}$$

α represents the margin, (v, e) denotes a matched image-text pair, and $(\hat{v}, \hat{e})$ refers to the hard negatives. $s(\cdot,\cdot)$ is the cosine similarity, and $[\cdot]_+$ indicates the hinge function $\max(\cdot, 0)$. The identification loss is also employed for training

$$L_c^o = CE(y, p_v) + CE(y, p_e) \tag{7}$$

where $CE(\cdot,\cdot)$ is cross entropy loss, y is identity label, p_v and p_e are the prediction probabilities. We perform bidirectional triplet ranking loss L_r^s and identification loss L_c^s at every granularity. The final training loss of student network is computed by

$$L_{stu} = \sum_{s \in \{o,1,2,3\}} (L_r^s + L_c^s) \tag{8}$$

Teacher Network. The teacher network employs a parameter-free dual attention module that utilizes cross-modality mixed features as guidance to determine the areas that require emphasis or restraint. The mixed features are derived from the average of image and text embeddings.

$$m_k^s = (v_k^s + e_k^s)/2 \tag{9}$$

Then, we perform a dot product between single modality features and mixed features to determine which ones to amplify or suppress

$$\begin{aligned} h_k^s &= v_k^s \cdot (1 + \tanh(v_k^s \cdot m_k^s))/2, \\ z_k^s &= e_k^s \cdot (1 + \tanh(e_k^s \cdot m_k^s))/2, \end{aligned} \tag{10}$$

Modality-shared semantics will have a larger response after the dot product. The generated attentive features h_k^s and z_k^s can highlight modality-shared information. We input the mixed features m_k^s, attentive image features h_k^s and attentive text features z_k^s into a shared CGAM module, the obtained mixed person representation is m, attentive image representation is h and attentive text representation is z. We adopts triple hard loss as

$$L_h = [\alpha - s(m,h) + s(m,\hat{h})]_+ + [\alpha - s(m,z) + s(m,\hat{z})]_+ \tag{11}$$

where (m, h) and (m, z) are the matched pairs, $(\hat{h}, \hat{z})$ are hard negatives. We introduce a cross-modal feature adaptation loss as

$$L_k = (\frac{1}{3d}\sum_i^{3d}(h_i)^2 - \frac{1}{3d}\sum_i^{3d}(z_i)^2)^2 \tag{12}$$

where i is feature channel. The final loss of teacher network is computed by

$$L_{tea} = L_h + L_k \tag{13}$$

IV. Experiment

A. Experiment Setting

We validate our method on three challenging text-to-image person retrieval datasets: CUHK-PEDES [1], ICFG-PEDES [3], and RSTPReid [2]. We resize all images to $384 \times 192 \times 3$ and normalize them with $1.0/256$. Word embeddings have a dimension of 400, while the embedding size for Bi-GRU is 512. he multi-head attention module consists of 6 heads, each with a feature dimension of 128. During training, image data augmentation is performed using random horizontal flipping, random crop with padding, and random erasing. Words are masked with a 0.15 probability. We use Adam as the optimizer, with an initial learning rate of 0.0002 and a weight decay of 5e-4. We set α equal to 0.2. The mini-batch size is 64, and the model is trained for 50 epochs. We evaluate the model using widely-used rank-k (k=1,5,10) matching accuracy and mean Average Precision (mAP).

TABLE I: Performance comparisons with state-of-the-art methods

Method	CUHK-PEDES				ICFG-PEDES				RSTPReid			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
ACSA(2022) [20]	63.56	81.40	87.70	-	-	-	-	-	48.40	71.85	81.45	-
SAF(2022) [21]	64.13	82.62	88.40	58.61	54.86	72.13	79.13	32.76	44.05	67.30	76.25	36.81
CAIBC(2022) [22]	64.43	82.87	88.37	-	-	-	-	-	47.35	69.55	79.00	-
C_2A_2(2022) [23]	64.82	83.54	89.77	-	-	-	-	-	51.55	76.75	85.15	-
LGUR(2022) [8]	65.25	83.12	89.00	-	59.02	75.32	81.56	-	-	-	-	-
IVT(2022) [24]	65.59	83.11	89.21	60.66	56.04	73.60	80.22	-	46.70	70.00	78.80	-
ASAMN(2023) [11]	65.66	84.53	90.21	-	57.09	76.33	82.84	-	-	-	-	-
LCR^{2}S(2023) [25]	67.36	84.19	89.62	59.24	57.93	76.08	82.40	38.21	54.95	76.65	84.70	40.92
UniPT(2023) [26]	68.50	84.67	90.38	-	60.09	76.19	82.46	-	51.85	74.85	82.85	-
UNIReID(2023) [27]	68.71	85.35	90.84	-	61.28	77.40	83.16	-	60.25	79.85	87.10	-
CFine(2023) [28]	69.57	85.93	91.15	-	60.83	76.55	82.42	-	50.55	72.50	81.60	-
TP-TPS(2023) [29]	70.16	86.10	90.98	66.32	60.64	75.97	81.76	**42.78**	50.65	72.45	81.20	43.11
VGSG(2023) [30]	71.38	86.75	91.86	67.91	63.05	78.43	84.36	-	-	-	-	-
TBPS-CLIP(2023) [31]	72.66	88.14	92.72	64.97	64.52	80.03	85.39	39.54	62.10	81.90	87.75	48.00
IRRA(2023) [9]	73.38	89.93	93.71	66.13	63.46	80.25	85.82	38.06	60.20	81.30	88.20	47.17
Baseline	65.83	84.46	90.52	60.78	54.21	74.81	81.79	31.92	53.14	79.55	86.55	43.62
OGEN	74.37	89.78	93.92	67.56	64.10	80.03	85.64	39.95	64.23	83.74	90.90	49.77
OGEN(RR)	**77.30**	**90.42**	**94.28**	**68.35**	**66.27**	**80.76**	**85.77**	41.86	**66.82**	**85.05**	**91.47**	**50.60**

TABLE II: Ablation study on each component of OGEN on three datasets. 'MGE' is multi-granularity embedding, 'CGAM' is cross-granularity aggregation module, 'DA' is dual attention module in teacher network.

Method		Student	Teacher		CUHK-PEDES			ICFG-PEDES			RSTPReid		
	MGE	CGAM	DA	CGAM	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
Baseline					65.83	84.46	90.52	54.21	74.81	81.79	53.14	79.55	86.55
+DA			✓		67.72	85.65	91.74	56.51	76.26	82.21	55.50	80.69	87.23
+MGE	✓				69.40	87.23	92.17	58.85	77.10	83.75	57.40	81.84	88.19
+MGE+Student	✓	✓			71.38	88.41	93.00	61.08	78.33	84.62	60.71	82.88	89.74
+MGE+DA	✓		✓		72.19	88.71	93.15	62.37	78.80	84.26	61.48	82.66	90.00
OGEN	✓	✓	✓	✓	74.37	89.78	93.92	64.10	80.03	85.64	64.23	83.74	90.90

TABLE III: Performance comparison of shared CGAM and no shared CGAM on CUHK-PEDES.

model	Rank-1	Rank-5	Rank-10
no shared CGAM	73.06	88.95	93.20
shared CGAM	74.37	89.78	93.92

B. Comparison with State-of-the-art Methods

In Tab. I, we present comparison results with state-of-the-art methods on three public benchmarks. The Baseline model employs global image-text embeddings, trained using bidirectional triplet ranking loss L_r and identification loss L_c. OGEN is our proposed omni-granularity embedding learning network. OGEN(RR) refers to the OGEN model with a re-ranking trick. We selected methods from the past two years for comparison. Most methods prior to 2023 used CNN as image feature extractors, while IVT and subsequent methods introduced ViT [24] as image feature extractors, and LCR^{2}S [25] employed ResNet for extracting image features.

Tab. I presents comparison results with state-of-the-art methods on three public benchmarks. Baseline model adopts global image-text embeddings (Granularity 1) with bidirectional triplet ranking loss L_r and identification loss L_c. OGEN is our proposed omni-granularity embedding network. OGEN(RR) refers to the OGEN model with a re-ranking trick. We selected methods from the past two years for comparison. Most methods prior to 2023 used CNN as image feature extractors, while IVT and subsequent methods introduced ViT [24] as image feature extractors, and LCR^{2}S [25] employed ResNet for extracting image features. CFine(2023) [28], TF-TPS [29], TBPS-CLIP [31] and IRRA [9] employ pre-trained CLIP models to extract visual and textual representations. Notably, CLIP-based methods achieve better performance, indicating that more robust pre-trained models can extract better person representations. Despite the success of existing methods, our OGEN outperforms all existing methods. Introducing the re-rank trick further elevates Rank-1 accuracies by 2%. OGEN consistently demonstrates the best performance for all metrics on all three benchmark datasets. Compared to the baseline, it improves Rank-1 accuracies by 8.54%, 9.89%, and 11.09%, respectively. This showcases the robustness and generalization of our proposed method.

C. Ablation Study

In this section, we analyze the effectiveness of each component in the OGEN framework.

Ablations on proposed components. We conduct a comprehensive empirical analysis to fully evaluate the impact of different components in CGAM. The Rank-1, Rank-5, and Rank-10 accuracies (%) are reported in Tab. II. The student network and teacher network share the same CGAM, with the DA being parameter-free. We have the following discussions: 1) OGEN aims to dynamically combine various granularity features to learn omni-granularity person representations. The effectiveness of multi-granularity embedding is demonstrated

through the experimental results of 'Baseline+MGE'. Simply adding MGE to the Baseline improves the Rank-1 accuracies by 3.57%, 4.64%, and 4.26% on the three datasets, respectively. These improvements highlight the importance of fine-grained representations extracted by MGE for text-to-image person retrieval. 2) Teacher network and student networks are weight sharing, with the teacher introducing an extra DA mechanism to emphasize modality-shared information. 'Baseline+DA' assesses the impact of DA. Simply adding DA to Baseline enhances Rank-1 accuracies by 1.89%, 2.29%, and 2.36% on the three datasets, respectively. When using DA in conjunction with MGE, model performance can also be improved. The findings demonstrate that unmatched contents may interfere with image-text matching. 3) We also eliminate the teacher network to assess the CGAM's ('Baseline+MGE+Student') contribution. In collaboration with MGE, CGAM enhances Rank-1 accuracies by 1.98%, 2.23%, and 3.31% on the three datasets, respectively. Without CGAM, all metrics decline. These results reveal that CGAM effectively aligns multi-granularity objects in a common space. It also indicates that the text-to-image person retrieval task is confronted by the omni-granularity matching issue. 4) To demonstrate the contribution of weight-sharing, we compare models with and without shared CGAM between teacher and student in Tab. III. The results reveal that sharing CGAM significantly enhances model performance. The performance improvements illustrate that weight-sharing aids the student network in learning modality-invariant person representations.

TABLE IV: Performance comparison of different number of local parts of each granularity on CUHK-PEDES dataset.

Grularity 2	Grularity 3	Rank-1	Rank-5	Rank-10
3	6	**74.37**	**89.78**	**93.92**
3	8	73.50	88.32	93.11
4	6	72.74	87.60	92.30
4	8	72.42	88.08	92.52

Different number of local parts. In Tab. IV, we compare the experimental results of varying the number of local parts on the CUHK-PEDES dataset. It can be seen that more parts do not always lead to better performance. A higher degree of granularity can disrupt the object structure, resulting in the model struggling to acquire high-level semantic information. The combination of '3,6' parts demonstrates the best performance.

TABLE V: Evaluation of cross-modal feature adaptation loss on CUHK-PEDES dataset. L_k(S) is model with feature adaptation on student network. L_k(T) is model with feature adaptation on teacher network.

model	Rank-1	Rank-5	Rank-10
w/o L_k	72.87	88.34	92.70
L_k(S)	73.24	88.65	93.31
L_k(T)	**74.37**	**89.78**	**93.92**

Cross-modal feature adaptation. We also evaluate the impact of cross-modal feature adaptation loss (L_k). Tab. V presents three settings: without L_k, applying L_k in the student network, and applying L_k in the teacher network. Notably, L_k(T) yields the superior performance. This is attributed to cross-modal feature adaptation enabling the teacher network to focus more on modality-shared features. The weight-shared CGAM can then transfer these emphasised features to the student network. As a result, OGEN can effectively learn modality-invariant person representations.

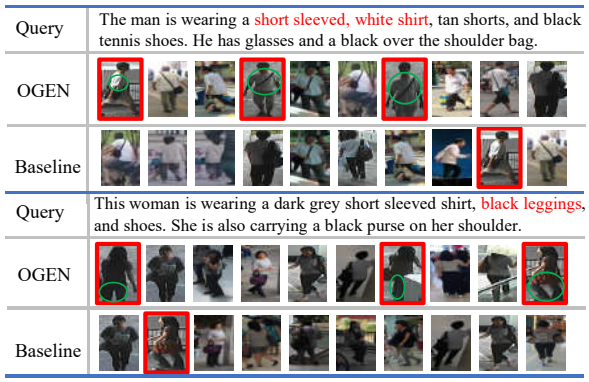


Fig. 3: Comparison of Rank-10 retrieved results on CUHK-PEDES between OGEN (the first row) and Baseline (the second row) for each text query. OGEN can dynamically match words with multi-granularity image regions.

Qualitative Results. Fig 3 illustrates the Rank-10 retrieval results of the Baseline and our proposed OGEN, images with red boxes are the right person. It clearly demonstrates that OGEN yields more accurate retrieval outcomes, whereas the Baseline miss some instances. This superiority is primarily attributed to our proposed CGAM (omni-granularity aggregation module), which efficiently learns to align words with multi-granularity image regions. Additionally, the target person in OGEN's results exhibits significant pose changes compared to the Baseline. This revelation showcases our OGEN's effectiveness in overriding intra-class variance, thereby proving that our approach significantly alleviates the omni-granularity matching issue.

V. Conclusion

In this article, we present OGEN to tackle the omni-granularity matching challenge. OGEN employs a novel cross-granularity aggregation module that dynamically integrates diverse granularity features, generating omni-granularity person representations within a shared embedding space. This enables words to align with multi-granularity image regions. OGEN further establishes a teacher-student knowledge transfer framework to teach the cross-granularity aggregation module and facilitate learning cross-modal person representations. By addressing the omni-granularity matching issue, OGEN sets a new state-of-the-art on three public datasets.

ACKNOWLEDGEMENT

This work was partially supported by the China Postdoctoral Science Foundation (2023M741305), the Fundamental Research Funds for the Central Universities (CCNU23XJ001) and the Open Project Program of Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industr, Wuyi University (FKLBDAITI202304).

REFERENCES

[1] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang, "Person search with natural language description," in *CVPR*, 2017.

[2] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua, "Dssl: Deep surroundings-person separation learning for text-based person retrieval," in *ACM MM*, 2021.

[3] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao, "Semantically self-aligned network for text-to-image part-aware person re-identification," *arXiv preprint arXiv:2107.12666*, 2021.

[4] Ying Zhang and Huchuan Lu, "Deep cross-modal projection learning for image-text matching," in *ECCV*, 2018.

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

[6] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li, "Text-based person search via multi-granularity embedding learning," in *IJCAI*, 2021.

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

[8] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding, "Learning granularity-unified representations for text-to-image person re-identification," in *ACM MM*, 2022.

[9] Ding Jiang and Mang Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *CVPR*, 2023.

[10] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan, "Pose-guided multi-granularity attention network for text-based person search," in *AAAI*, 2020.

[11] Kai Niu, Tao Huang, Linjiang Huang, Liang Wang, and Yanning Zhang, "Improving inconspicuous attributes modeling for person search by language," *IEEE TIP*, 2023.

[12] Kajal Kansal, A.V. Subramanyam, Zheng Wang, and Shinichi Satoh, "Hierarchical attention image-text alignment network for person re-identification," in *ICME Workshops*, 2021.

[13] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang, "Omni-scale feature learning for person re-identification," in *ICCV*, 2019.

[14] Zhaoxiang Zhang, Chuanchen Luo, Haiping Wu, Yuntao Chen, Naiyan Wang, and Chunfeng Song, "From individual to whole: reducing intra-class variance by feature aggregation," *IJCV*, 2022.

[15] Jianyang Gu, Kai Wang, Hao Luo, Chen Chen, Wei Jiang, Yuqiang Fang, Shanghang Zhang, Yang You, and Jian Zhao, "Msinet: Twins contrastive search of multi-scale interaction for object reid," in *CVPR*, 2023.

[16] Kaixiang Chen, Tiantian Gong, and Liyan Zhang, "Multi-scale query-adaptive convolution for generalizable person re-identification," in *ICME*, 2023.

[17] Keyu Wen, Jin Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, and Jie Shao, "Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation," in *ICCV*, 2021.

[18] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *CVPR*, 2022.

[19] Qi Zhao, Shuchang Lyu, Lijiang Chen, Binghao Liu, Ting-Bing Xu, Guangliang Cheng, and Wenquan Feng, "Learn by oneself: Exploiting weight-sharing potential in knowledge distillation guided ensemble network," *IEEE TCSVT*, 2023.

[20] Zhong Ji, Junhua Hu, Deyin Liu, Lin Yuanbo Wu, and Ye Zhao, "Asymmetric cross-scale alignment for text-based person search," *IEEE TMM*, 2022.

[21] Shiping Li, Min Cao, and Min Zhang, "Learning semantic-aligned feature representation for text-based person search," in *ICASSP*, 2022.

[22] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li, "Caibc: Capturing all-round information beyond color for text-based person retrieval," in *ACM MM*, 2022.

[23] Kai Niu, Linjiang Huang, Yan Huang, Peng Wang, Liang Wang, and Yanning Zhang, "Cross-modal co-occurrence attributes alignments for person search by language," in *ACM MM*, 2022.

[24] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang, "See finer, see more: Implicit modality alignment for text-based person retrieval," in *ECCV*, 2022.

[25] Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang, "Learning comprehensive representations with richer self for text-to-image person re-identification," in *ACM MM*, 2023.

[26] Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang, "Unified pre-training with pseudo texts for text-to-image person re-identification," in *ICCV*, 2023.

[27] Cuiqun Chen, Mang Ye, and Ding Jiang, "Towards modality-agnostic person re-identification with descriptive query," in *CVPR*, 2023.

[28] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang, "Clip-driven fine-grained text-image person re-identification," *IEEE TIP*, 2023.

[29] Guanshuo Wang, Fufu Yu, Junjie Li, Qiong Jia, and Shouhong Ding, "Exploiting the textual potential from vision-language pre-training for text-based person search," *arXiv preprint arXiv:2303.04497*, 2023.

[30] Shuting He, Hao Luo, Wei Jiang, Xudong Jiang, and Henghui Ding, "Vgsg: Vision-guided semantic-group network for text-based person search," *IEEE TIP*, 2023.

[31] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang, "An empirical study of clip for text-based person search," *arXiv preprint arXiv:2308.10045*, 2023.