Cross-Modal Feature Fusion-Based Knowledge Transfer for Text-Based Person Search

Kaiyang You, Wenjing Chen, Chengji Wang, Hao Sun, *Member, IEEE*, and Wei Xie, *Member, IEEE*

***Abstract*—Text-based person search aims to retrieve corresponding images of person from a large gallery based on text descriptions. Existing methods strive to bridge the modality gap between images and texts and have made promising progress. However, these approaches disregard the knowledge imbalance between images and texts caused by the reporting bias. To resolve this issue, we present a cross-modal feature fusion-based knowledge transfer network to balance identity information between images and texts. First, we design an identity information emphasis module to enhance person-relevant information and suppress person-irrelevant information. Second, we design an intermediate modal-guided knowledge transfer module to balance the knowledge between images and texts. Experimental results on CUHK-PEDES, ICFG-PEDE, and RSTPReid datasets demonstrate that our method achieves state-of-the-art performance.**

***Index Terms*—Text-based person search, knowledge imbalance, knowledge transfer, cross-modal fusion.**

I. Introduction

TEXT-based Person Search (TPS) aims to retrieve corresponding person images from a large gallery based on a textual description [1]. This task is becoming a research hotspot because of its potential applications in missing person searching, finding suspects [2], [3], [4]. As a cross-modal retrieval task, the modality heterogeneity between vision and language makes TPS a challenging task [5].

The key to TPS lies in cross-modal alignment between images and text descriptions. Early methods [1], [6], [7], [8], [9], [10] focused on aligning global features of images and texts. However, these approaches can not effectively explore the discriminative local details of images and texts, which are key clues for TPS. Some subsequent methods [11], [11], [12], [13], [14], [15] focused on designing different components (e.g. pose estimation [12], attribute segmentation [13], and image/text splitting [11]) to explicitly extract local features. Then, these methods establish correspondences between local features of images and texts for feature alignment. However, these methods require additional components to extract local features, increasing computational costs. In addition, it is challenging to accurately extract local features for each person, and wrong local features will cause interference. The latest method IRRA [2], LSPM [16], MUM [17] and BiLMa [18] leverages Masked Language Modeling (MLM) to implicitly establish local correspondences between images and texts without using external components, improving retrieval performance.

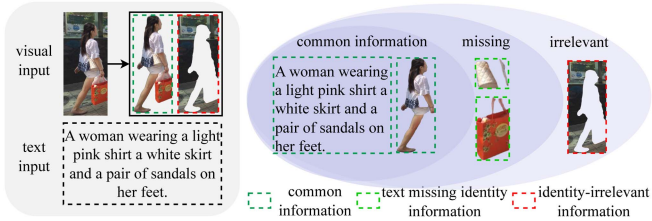

Fig. 1. The illustration of knowledge imbalance between images and texts in this task. The text only contains part of the information in the image. The image and text are from the CUHK-PEDES dataset [1].

Although recent methods have achieved gratifying progress, these methods treat images and texts equally, without considering the knowledge imbalance between images and texts caused by the reporting bias [4]. Texts are the subjective portrayals of the target character by human respondents, may contain only partial information about the person to be retrieved. As shown in Fig. 1, in this sample pair, the text descriptions lack information such as a white bag and a red shopping bag, which are important for proper TPS. In order to alleviate the problem of knowledge imbalance, existing methods [4], [8] usually adopt a knowledge transfer strategy to transfer rich knowledge in images to texts. However, these methods still have two problems. Firstly, the image usually contains person-irrelevant information, such as background and occlusion. Transferring person-irrelevant information may weaken the discrimination of text features. Secondly, since there is huge modal gaps [19], [20], [21], [22] between images and texts, direct cross-modal knowledge transferring from images to texts is challenging [23].

In this letter, we propose a Cross-modal Feature Fusion-based Knowledge Transfer (CFFKT), which can effectively alleviate the knowledge imbalance for TPS. First, we propose an Identity Information Emphasis (IIE) module that exploits the commonalities among various person images to enhance

Received 25 May 2024; revised 14 August 2024; accepted 18 August 2024. Date of publication 23 August 2024; date of current version 5 September 2024. This work was supported in part by National Natural Science Foundation of China under Grant 62201222 and Grant 62377026, in part by the Knowledge Innovation Program of Wuhan-Shuguang Project under Grant 2023010201020382 and Grant 2023010201020377, and in part by the Fundamental Research Funds for the Central Universities under Grant CCNU22QN014, Grant CCNU24ai011, and Grant CCNU24JCPT027. The associate editor coordinating the review of this article and approving it for publication was Dr. Mingye Ju. *(Corresponding authors: Hao Sun; Wei Xie.)*

Kaiyang You, Chengji Wang, Hao Sun, and Wei Xie are with the Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, the School of Computer Science, and the National Language Resources Monitoring and Research Center for Network Media, Central China Normal University, Wuhan 430079, China (e-mail: ykyou@mails.ccnu.edu.cn; wcj@ccnu.edu.cn; haosun@ccnu.edu.cn; xw@mail.ccnu.edu.cn).

Wenjing Chen is with the School of Computer Science, Hubei University of Technology, Wuhan 430068, China (e-mail: chenwenjing@hbut.edu.cn).

Digital Object Identifier 10.1109/LSP.2024.3449222

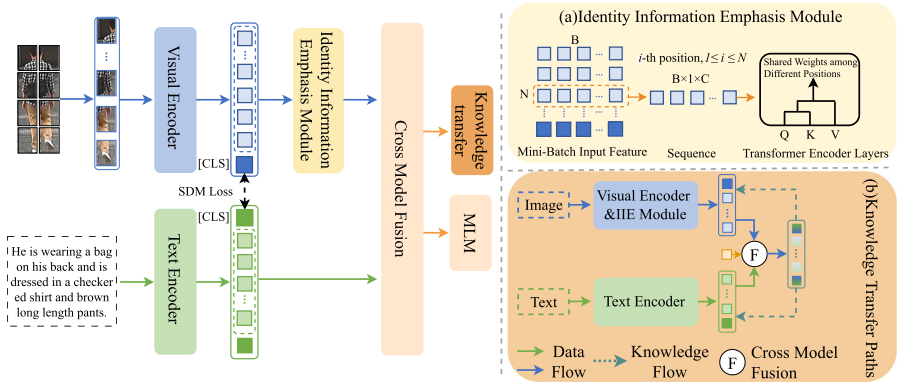

Fig. 2. Overall architecture of the proposed CFFKT.

person-relevant information and suppress person-irrelevant information. Second, we propose an Intermediate Modal-guided Knowledge Transfer (IMKT) module to balance the knowledge between images and texts, which includes a Cross-Modal Fusion (CMF) component and a Knowledge Transfer (KT) component. CMF aims to bridge the inherent modal gaps by fusing the image and text features to obtain the intermediate-modal features, which contain the knowledge from both images and texts [24]. The goal of KT is to transfer rich knowledge from intermediate-modal feature to text and image, respectively, which can reduce the difficulty of direct cross-modal knowledge transferring. Here, intermediate-modal-to-text knowledge transferring is used to transfer image knowledge to text features to alleviate knowledge imbalance. Intermediate-modal-to-image knowledge transferring aims to utilize the text information to further suppress person-irrelevant information contained in image features. Extensive experiments on CUHK-PEDES [1], ICFG-PEDES [25] and RSTPReid [26] evaluate the superior performance of CFFKT.

II. Proposed Method

In this section, we will introduce our CFFKT. The overall structure of CFFKT is shown in Fig. 2, and the specifics are introduced in subsequent subsections.

A. Feature Extraction

Recent methods [2], [27] have demonstrated that the pre-trained CLIP model can provide strong potential cross-modal alignment capabilities for feature extractors. Therefore, we employ the visual and text encoder from CLIP as feature extractors. Giving an image and its corresponding text, we input the image to the visual encoder to obtain a sequence of visual features $\{v_{\text{cls}}, v_1, \ldots, v_N\}$, where $v_i \in \mathbb{R}^C$, N represents the count of flattened patches, and C is the feature dimension. Likewise, we obtain a sequence of text features $\{t_{\text{sos}}, t_1, \ldots, t_L, t_{\text{eos}}\}$ by feeding the query text into the text encoder, where $t_i \in \mathbb{R}^C$. t_{sos} and t_{eos} represent the beginning and end of the text sequence, respectively.

B. The Identity Information Emphasis Module

In TPS, the various parts of the person body are evenly arranged in the image. The commonalities of the regular arrangement among persons in different images can be used to emphasize person-relevant features and suppress person-irrelevant features [28]. The proposed IIE module utilizes a stack of multiple encoder layers in Transformer [29] to model the commonalities of person-image relationships. The input of the IIE module will first be reshaped to enable the transformer layer to handle the batch dimensions of the input data. By doing this, the self-attention mechanism in the transformer layer becomes the batch attention mechanism between different samples of IIE. The batch attention mechanism can mine common knowledge among batch data and discard category-irrelevant information [28], [30]. Our images are all pedestrians, and the human bodies are evenly arranged in the image, so we can enhance the category features related to people in the image and suppress the features unrelated to people. As shown in the Fig. 2(a), among the features of different images in a mini-batch, we regard the patch features at the ith $(1 \leq i \leq N)$ position of different images as a sequence. Hence, we have N sequences, and each sequence length is B. All the above sequences are then fed into the transformer encoder layers to learn the emphasized image features, where person-relevant features are emphasized and person-irrelevant features are suppressed [28].

C. The Intermediate Modal-Guided Knowledge Transfer Module

Although the IIE module suppresses non-person-related information in images, there is an inherent modal gap between images and texts. If we directly perform cross-modal knowledge transfer, it will affect the performance of the transfer. Therefore, we propose an IMKT module that includes a CMF component and a KT component. We utilize CMF to bridge the inherent modal gap and obtain intermediate modal features containing image and text knowledge. The goal of KT is to transfer rich knowledge from Intermediate modal features are transferred to text and images respectively, which can reduce the difficulty of direct cross-modal knowledge transfer and achieve effective knowledge transfer.

TABLE I
COMPARING WITH THE STATE-OF-THE-ART PERFORMANCE ON THE CUHK-PEDES, ICFG-PEDES AND RSTPREID

Method	Ref	CUHK-PEDES				ICFG-PEDES				RSTPReid			
		R@1	R@10	mAP	mINP	R@1	R@10	mAP	mINP	R@1	R@10	mAP	mINP
RKT [4]	TMM23	61.48	87.28	-	-	-	-	-	-	-	-	-	-
TGDA [3]	TCSVT23	64.64	89.34	58.64	-	57.26	81.8	32.27	-	48.35	80.3	37.96	-
MANet [31]	TNNLS23	65.64	88.78	-	-	59.44	82.75	-	-	-	-	-	-
LCR^{2}S [32]	MM23	67.36	89.62	59.24	-	57.93	82.40	38.21	-	54.95	84.70	40.92	-
UniPT [33]	ICCV23	68.50	90.38	-	-	60.09	82.46	-	-	51.85	82.85	-	-
TMA [34]	CVPR23	68.71	90.84	44.28	-	61.28	83.16	-	-	60.25	87.10	-	-
CTL [35]	TCSVT23	69.47	92.13	60.56	-	57.69	82.67	36.07	-	-	-	-	-
VGSG [36]	TIP24	71.38	91.86	-	-	63.05	84.36	-	-	-	-	-	-
IRRA [2]	CVPR23	73.38	93.71	66.13	50.24	63.46	**85.82**	38.06	7.93	60.20	88.20	47.17	25.28
FedSH [37]	TMM24	57.48	86.00	-	-	51.07	77.57	-	-	-	-	-	-
EAIBC [38]	TNNLS24	64.96	88.42	-	-	58.95	81.72	-	-	49.83	79.85	-	-
Baseline	-	69.81	91.95	63.97	49.34	59.63	81.79	37.14	7.56	58.50	86.85	46.54	24.51
OURS	-	**73.91**	**93.91**	**66.56**	**51.10**	**64.22**	85.54	**39.28**	**8.54**	**60.95**	**88.55**	**47.67**	**25.66**

The bold values indicate the best result.

TABLE II
ABLATION STUDY ON EACH COMPONENT OF CFFKT ON CUHK-PEDES, ICFG-PEDES AND RSTPREID

No.	Methods	Components		CUHK-PEDES			ICFG-PEDES			RSTPReid		
		IIE	IMKT	R@1	R@10	mAP	R@1	R@10	mAP	R@1	R@10	mAP
0	Baseline			69.81	91.95	63.97	59.63	81.79	37.14	58.50	86.85	46.54
1	CMKT			69.98	92.33	62.56	59.37	83.46	33.78	58.35	87.35	45.59
2	+IIE	✓		70.60	92.30	62.86	61.59	84.31	36.21	59.05	88.45	46.59
3	+IMKT		✓	72.56	93.34	65.43	63.33	85.33	39.45	59.45	87.70	46.82
4	+IMKT+IIE	✓	✓	73.32	93.45	65.90	63.58	**85.63**	38.92	60.00	87.85	47.49
5	CFFKT	✓	✓	**73.91**	**93.91**	**66.56**	**64.22**	85.54	**39.28**	**60.95**	**88.55**	**47.67**

The bold values indicate the best result.

Cross-Modal Fusion Component: As depicted in Fig. 2, we concat the emphasised image features $\{v_{\texttt{cls}}, v_1, \ldots, v_N\}$ and text features $\{t_{\texttt{sos}}, t_1, \ldots, t_L, t_{\texttt{eos}}\}$ as the input of the CMF component, and add a randomly initialized learnable token [UNI], which is utilized to learn intermediate-modal feature. Specifically, the CMF component is composed of 9-layer transformer encoders with 8 heads and 512 dimensional width. The output of CMF is $\{f_{v,\texttt{cls}}, f_{v,1}, \ldots, f_{v,N}, f_{\texttt{uni}}, f_{t,\texttt{sos}}, f_{t,1}, \ldots, f_{t,l}, f_{t,\texttt{eos}}\}$. The fused multi-modal features $f_{\texttt{uni}}$ can be used as an intermediate modality between images and texts [24]. Using intermediate-modal features $f_{\texttt{uni}}$ for knowledge transfer can reduce the difficulty of knowledge transferring.

Knowledge Transfer Component: The intermediate-modal feature $f_{\texttt{uni}}$ contain rich identity knowledge. Transferring the knowledge of the intermediate modality to the text modality can supplement the identity information of the text features to alleviate knowledge imbalance [24], [39], [40]. Transferring knowledge from intermediate modality to image modality can utilize text information to further suppress person-irrelevant information contained in image features.

At the same time, considering the heterogeneity between different modals, we employ InfoNCE loss [41] to perform contrastive learning in the potential feature space to conduct the intermediate-modal-to-text knowledge transferring and the intermediate-modal-to-image knowledge transferring. The knowledge transfer loss between j-th pair of intermediate-modal features $f_{\texttt{uni},j}$ and text features $t_{\texttt{eos},j}$ in a mini-batch is as follows:

$$\mathcal{L}_{i\to t} = -\log\frac{\exp(\Phi(f_{\texttt{uni},j}, t_{\texttt{eos},j}/)\tau)}{\sum_{h=1}^{B}\exp(\Phi(f_{\texttt{uni},h}, t_{\texttt{eos},h}/)\tau)}, \quad (1)$$

where τ is the temperature coefficient, Φ is a cosine similarity. This formula represents the probability of assigning the intermediate-modal feature $f_{\texttt{uni},j}$ to its paired text feature $t_{\texttt{eos},j}$ against the whole mini-batch of text representation $\{t_{\texttt{eos},j}\}_{j=1}^{B}$, where B is the batch size.

To constrain the visual encoder in the feature extractor to learn person-relevant features, we directly transfer the knowledge of intermediate-modal features to initial image features $\{v_{\texttt{cls}}, v_1, \ldots, v_N\}$. Similarly, the loss of intermediate-modal-to-image knowledge transferring is as follows:

$$\mathcal{L}_{i\to v} = -\log\frac{\exp(\Phi(f_{\texttt{uni},j}, v_{\texttt{cls},j}/)\tau)}{\sum_{h=1}^{B}\exp(\Phi(f_{\texttt{uni},h}, v_{\texttt{cls},h}/)\tau)}, \quad (2)$$

During the training phase, the proposed CFFKT contains four objective functions. Following IRRA [2], we adopt SDM loss $\mathcal{L}_{\texttt{sdm}}$ and MLM loss $\mathcal{L}_{\texttt{mlm}}$ to diminish the gap between image and text modalities. And in order to stimulate the backbone model to extract discriminative image and text features, we introduce ID loss $\mathcal{L}_{\texttt{ID}}$ [13]. In summary, CFFKT is trained using an end-to-end approach, and the final optimization objective function $\mathcal{L}$ is as:

$$\mathcal{L} = \mathcal{L}_{i\to t} + \mathcal{L}_{i\to v} + \mathcal{L}_{\texttt{mlm}} + \mathcal{L}_{\texttt{sdm}} + \mathcal{L}_{\texttt{ID}}. \quad (3)$$

In the inference stage, we directly use a visual feature extractor and a text feature extractor to extract features, and exploit the similarity between the global features of images and texts for person search without introducing extra computation costs.

III. EXPERIMENTAL RESULT

A. Datasets and Experiment Details

We perform experiments on three TPS benchmark datasets that are subject to reporting bias [4]. CUHK-PEDES [1] is the pioneering dataset specifically crafted for TPS, encompassing

80,412 text and 40,206 images of 13,003 individuals. ICFG-PEDES [25] is a new and more challenging TPS dataset, which encompasses 54,522 pedestrian images of 4,102 pedestrians. RSTPReid [26] is a real-world TPS dataset, which includes 20,505 images of 4,101 individuals.

In experiments, we use the CLIP-ViT-B/16 as the visual encoder, the CLIP text transformer as the text encoder. The weights of IIE and CMF are randomly initialized. We resize the size of images to 384×128 and limit the text token sequence length L to 77. While training, we adopt randomly erasing, horizontal flipping and random cropping as data augmentations. Our method is trained for 80 epochs using the Adam optimizer with an initial learning rate of 1×10^{-5} and cosine learning rate decay. We employ the R@k metrics (k=1,10), mAP [42] and mINP [2] as the evaluation metrics.

B. Comparison With State-of-The-Art Methods

We compared with other state-of-the-art methods on CUHK-PEDES [1], ICFG-PEDES [25], and RSTPReid [26] datasets. The experimental results are displayed in Table I. The Baseline in Tables I and II both represent CLIP-ViT-B/16 model fine-tuned with ID loss [13] and SDM loss [2]. As shown in Table I, our CFFKT achieved the best results on the three datasets. In these comparison methods, TGDA [3], MANet [31], UniPT [33], CTL [35] and IRRA [2] attaches importance to mining the fine-grained relationship between images and texts. However, these approaches regard images and texts alike, ignoring the identity knowledge imbalance between images and texts, which may limit their performance.

Furthermore, RKT [4] adopts direct image-to-text knowledge transfer to alleviate the knowledge imbalance between images and texts, which may not be optimal due to the modal heterogeneity between image modalities and text modalities. In contrast, our CFFKT utilizes the IMKT module to balance the knowledge between images and texts, reducing the difficulty of direct cross-modal knowledge transfer, thereby improving the performance of person search. To adequately illustrate the effect of different components on CFFKT, we implement a comprehensive empirical analysis on CUHK-PEDES [1], ICFG-PEDES [25] and RSTPReid [26] datasets. The accuracy (%) of R@1, R@10 and mAP is shown in Table II.

C. Ablation Study

Ablation of IIE Module: The IIE module significantly improves performance by using batch attention to model the relationship between people and images. Comparing results in No.0 vs. No.2, it is evident that enhancing person-relevant information and suppressing person-irrelevant information largely enhances the performance on R@1, obtaining improvement by 0.79%, 1.96%, and 0.55%. To additionally verify the effectiveness of IIE, we performed heatmap visualization experiments on the CUHK-PEDES dataset. As depicted in Fig. 3, it is evident that the attention of the network is more focused on people after IIE is added to the model, which shows that IIE is effective in emphasizing people-related features.

Ablation for IMKT Module: The IMKT module plays a crucial role in our approach to alleviate the imbalance of image and text identity information. Comparing the results of No.1 and No.3, we can notice that using the proposed cross-modal fusion knowledge transfer instead of the commonly used direct cross-modal knowledge transfer (CMKT), the R@1 of the three datasets increased by 2.58%, 3.96 % and 1.1%, respectively. Moreover, CMKT does not improve the performance of TPS but leads to partial performance degradation (No.0 *vs* No.1). Compared with CMKT, our use of IMKT can better alleviate the knowledge imbalance between images and texts and improve search accuracy. The introduction of the MLM loss (No.5 *vs* No.4) into the IMKT module can also effectively improve the performance of person search because MLM can help promote fine-grained feature alignment of the image-text model and bridge the gaps between modalities.

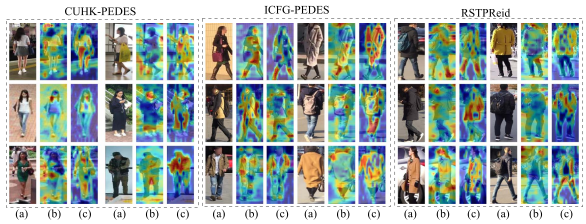

Fig. 3. The heat map visualization results on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets are presented as follows: columns (a) represent the source images, columns (b) show the heat maps generated after applying the baseline method, columns (c) columns show the heat maps obtained by our approach.

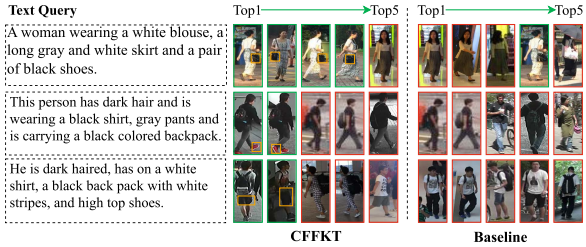

Fig. 4. Qualitative text-to-image retrieval results of CFFKT and Baseline on CUHK-PEDES. The green boxes indicate the correct matches, and the images in the red boxes are the wrong matches.

Qualitative Results: Fig. 4 presents a comparison of the top 5 retrieved results between the Baseline model and our proposed CFFKT. As depicted in the illustration, CFFKT obtains improved retrieval precision and succeeds in obtaining accurate results where the Baseline falls short. This is mainly due to the IMKT module we designed, which can still retrieve the correct results when the text description part is missing. The orange-highlighted box in Fig. 4 demonstrates this.

IV. Conclusion

In this work, we propose a cross-modal fusion knowledge transfer model for TPS. We propose the identity information emphasis module and the intermediate modal knowledge transfer module to alleviate the identity knowledge imbalance existing between images and texts. Extensive experiments on CUHK-PEDES, ICFG-PEDE and RSTPReid datasets confirm the efficacy and excellence of our method. In the future, we will study lightweight networks to improve retrieval efficiency.

REFERENCES

[1] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5187–5196.

[2] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2787–2797.

[3] L. Gao, K. Niu, B. Jiao, P. Wang, and Y. Zhang, "Addressing information inequality for text-based person search via pedestrian-centric visual denoising and bias-aware alignments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7884–7899, Dec. 2023.

[4] Z. Wu, B. Ma, H. Chang, and S. Shan, "Refined knowledge transfer for language-based person search," *IEEE Trans. Multimedia*, vol. 25, pp. 9315–9329, 2023.

[5] J. Zuo et al., "Ufinebench: Towards text-based person retrieval with ultra-fine granularity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 22010–22019.

[6] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Adversarial representation learning for text-to-image matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5813–5823.

[7] Z. Wang et al., "CAIBC: Capturing all-round information beyond color for text-based person retrieval," in *Proc. ACM Int. Conf. Multimedia.*, 2022, pp. 5314–5322.

[8] Y. Chen, R. Huang, H. Chang, C. Tan, T. Xue, and B. Ma, "Cross-modal knowledge adaptation for language-based person search," *IEEE Trans. Image Process.*, vol. 30, pp. 4057–4069, 2021.

[9] Z. Wang et al., "Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold," in *Proc. ACM Int. Conf. Multimedia.*, 2022, pp. 1984–1992.

[10] A. Farooq, M. Awais, J. Kittler, and S. S. Khalid, "AXM-Net: Implicit cross-modal feature alignment for person re-identification," in *Proc. Conf. Innov. Appl. Artif. Intell.*, 2022, pp. 4477–4485.

[11] Z. Ji, J. Hu, D. Liu, L. Y. Wu, and Y. Zhao, "Asymmetric cross-scale alignment for text-based person search," *IEEE Trans. Multimedia*, vol. 25, pp. 7699–7709, 2023.

[12] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-guided multi-granularity attention network for text-based person search," in *Proc. Innov. Appl. Artif. Intell. Conf.*, 2020, pp. 11189–11196.

[13] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "ViTAA: Visual-textual attributes alignment in person search by natural language," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2020, pp. 402–420.

[14] W. Suo et al., "A simple and robust correlation filtering method for text-based person search," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2022, pp. 726–742.

[15] C. Gao et al., "Conditional feature learning based transformer for text-based person search," *IEEE Trans. Image Process.*, vol. 31, pp. 6097–6108, 2022.

[16] J. Li, M. Jiang, J. Kong, X. Tao, and X. Luo, "Learning semantic polymorphic mapping for text-based person retrieval," *IEEE Trans. Multimedia*, early access, Jun. 05, 2024, doi: 10.1109/TMM.2024.3410129.

[17] Z. Zhao, B. Liu, Y. Lu, Q. Chu, and N. Yu, "Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 7534–7542.

[18] T. Fujii and S. Tarashima, "BiLMa: Bidirectional local-matching for text-based person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2023, pp. 2778–2782.

[19] A. Chaudhuri, M. Mancini, Y. Chen, Z. Akata, and A. Dutta, "Cross-modal fusion distillation for fine-grained sketch-based image retrieval," in *Proc. Brit. Mach. Vis. Conf.*, 2022. [Online]. Available: https://bmvc2022.mpi-inf.mpg.de/499/

[20] Z. Xue, Z. Gao, S. Ren, and H. Zhao, "The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation," in *Proc. Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=w0QXrZ3N-s

[21] H. Sun, M. Zhou, W. Chen, and W. Xie, "Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, pp. 4998–5007.

[22] M. Meng, J. Sun, J. Liu, J. Yu, and J. Wu, "Semantic disentanglement adversarial hashing for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1914–1926, Mar. 2024.

[23] S. Ren, Y. Du, J. Lv, G. Han, and S. He, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13320–13328.

[24] J. Huang, Y. Li, J. Feng, X. Wu, X. Sun, and R. Ji, "Clover: Towards a unified video-language alignment and fusion model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14856–14866.

[25] Z. Ding, C. Ding, Z. Shao, and D. Tao, "Semantically self-aligned network for text-to-image part-aware person re-identification," 2021, *arXiv:2107.12666*.

[26] A. Zhu et al., "DSSL: Deep surroundings-person separation learning for text-based person retrieval," in *Proc. ACM Multimedia Conf.*, 2021, pp. 209–217.

[27] M. Zhou, W. Chen, H. Sun, and W. Xie, "Cross-modal multiscale difference-aware network for joint moment retrieval and highlight detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 8416–8420.

[28] Z. Hou, B. Yu, and D. Tao, "Batchformer: Learning to explore sample relationships for robust representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7246–7256.

[29] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[30] Z. Hou, B. Yu, C. Wang, Y. Zhan, and D. Tao, "Batchformerv2: Exploring sample relationships for dense representation learning," 2022, *arXiv:2204.01254*.

[31] S. Yan, H. Tang, L. Zhang, and J. Tang, "Image-specific information suppression and implicit local alignment for text-based person search," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 15, 2023, doi: 10.1109/TNNLS.2023.3310118.

[32] S. Yan, N. Dong, J. Liu, L. Zhang, and J. Tang, "Learning comprehensive representations with richer self for text-to-image person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 6202–6211.

[33] Z. Shao, X. Zhang, C. Ding, J. Wang, and J. Wang, "Unified pre-training with pseudo texts for text-to-image person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11140–11150.

[34] C. Chen, M. Ye, and D. Jiang, "Towards modality-agnostic person re-identification with descriptive query," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15128–15137.

[35] H. Wu, W. Chen, Z. Liu, T. Chen, Z. Chen, and L. Lin, "Contrastive transformer learning with proximity data generation for text-based person search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7005–7016, Aug. 2024.

[36] S. He, H. Luo, W. Jiang, X. Jiang, and H. Ding, "VGSG: Vision-guided semantic-group network for text-based person search," *IEEE Trans. Image Process.*, vol. 33, pp. 163–176, 2024.

[37] W. Ma et al., "FedSH: Towards privacy-preserving text-based person re-identification," *IEEE Trans. Multimedia*, vol. 26, pp. 5065–5077, 2024.

[38] A. Zhu et al., "Improving text-based person retrieval by excavating all-round information beyond color," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 28, 2024, doi: 10.1109/TNNLS.2024.3368217.

[39] L. Zhang, L. Song, B. Du, and Y. Zhang, "Nonlocal low-rank tensor completion for visual data," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 673–685, Feb. 2021.

[40] Y. Ge et al., "Bridging video-text retrieval with multiple choice questions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16146–16155.

[41] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkgpBJrtvS

[42] X. Han, S. He, L. Zhang, and T. Xiang, "Text-based person search with limited data," in *Proc. Brit. Mach. Vis. Conf.*, 2021. [Online]. Available: https://www.bmvc2021-virtualconference.com/conference/papers/paper_0044.html