Adaptive Uncertainty-Based Learning for Text-Based Person Retrieval

Shenshen Li, Chen He, Xing Xu*, Fumin Shen, Yang Yang, Heng Tao Shen

School of Computer Science and Engineering and Center for Future Media,
University of Electronic Science and Technology of China, China
lishenshen727@gmail.com, chen_he1219@outlook.com, xing.xu@uestc.edu.cn, fumin.shen@gmail.com,
yang.yang@uestc.edu.cn, shenhengtao@hotmail.com

Abstract

Text-based person retrieval aims at retrieving a specific pedestrian image from a gallery based on textual descriptions. The primary challenge is how to overcome the inherent heterogeneous modality gap in the situation of significant intra-class variation and minimal inter-class variation. Existing approaches commonly employ vision-language pre-training or attention mechanisms to learn appropriate cross-modal alignments from noise inputs. Despite commendable progress, current methods inevitably suffer from two defects: 1) Matching ambiguity, which mainly derives from unreliable matching pairs; 2) One-sided cross-modal alignments, stemming from the absence of exploring one-to-many correspondence, *i.e.*, coarse-grained semantic alignment. These critical issues significantly deteriorate retrieval performance. To this end, we propose a novel framework termed *Adaptive Uncertainty-based Learning (AUL)* for text-based person retrieval from the uncertainty perspective. Specifically, our AUL framework consists of three key components: 1) Uncertainty-aware Matching Filtration that leverages Subjective Logic to effectively mitigate the disturbance of unreliable matching pairs and select high-confidence cross-modal matches for training; 2) Uncertainty-based Alignment Refinement, which not only simulates coarse-grained alignments by constructing uncertainty representations but also performs progressive learning to incorporate coarse- and fine-grained alignments properly; 3) Cross-modal Masked Modeling that aims at exploring more comprehensive relations between vision and language. Extensive experiments demonstrate that our AUL method consistently achieves state-of-the-art performance on three benchmark datasets in supervised, weakly supervised, and domain generalization settings. Our code is available at https://github.com/CFM-MSG/Code-AUL.

Introduction

The text-based person retrieval task (Li et al. 2017; Ding et al. 2021) aims at locating the specific pedestrian image from a collection of candidates with a provided textual description query. Compared with the conventional image-based (Specker, Cormier, and Beyerer 2023) or video-based person retrieval (Hou et al. 2021), the query of text-based person retrieval provides a readily accessible and intuitive means for describing the attributes of the target person, making it a popular and active area of research. However, owing to the substantial intra-class variation and minimal inter-class variation, the text-based person retrieval task faces increased challenges in overcoming the inherent heterogeneous modality gap (Jiang et al. 2022; Li et al. 2023a), which considerably hampers the overall retrieval performance.

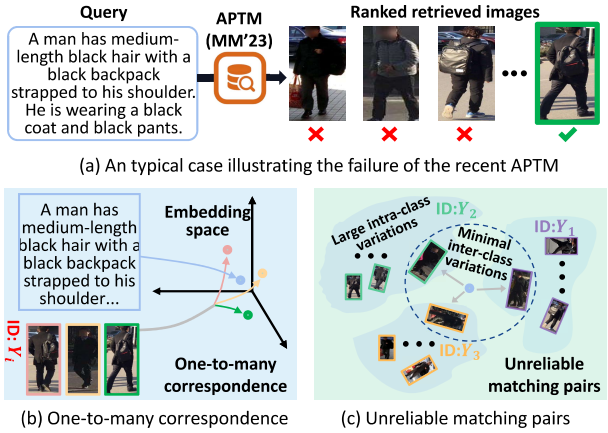

Figure 1. Illustrative examples of existing problems: (a) A representative failure case of the latest method APTM. (b) The presence of one-to-many correspondence is evident. (c) Unreliable matching pairs that stem from large intra-class variation and minimal inter-class variation.

To tackle the above problem, previous methods mainly focus on cross-modal alignments by the utilization of supplementary information (Zhu et al. 2021) or the incorporation of diverse attention mechanisms (Suo et al. 2022). Moreover, recent approaches (Jiang and Ye 2023) have been influenced by the capabilities of vision-language pre-training to enhance representation learning, which can effectively characterize the association between vision and language.

As shown in Figure 1(a), while the mentioned methods have made advancements, they still suffer from matching ambiguity and one-sided cross-modal alignments, thus leading to decreased performance and limited generalization. Such two problems can be attributed to the following aspects: 1) Absence of one-to-many correspondence, that stems from the limitation of considering only one-to-

*Corresponding Author

one matching. As shown in Figure 1(b), there indeed exists the one-to-many correspondence between language and vision. Specifically, the visual data could thoroughly capture all the objects yet lack context as in the corresponding text, and the language could not fully describe every detail of a scene based on the human-annotated caption. Such instinctive nature leads to the necessity of exploring one-to-many correspondence between vision and language. 2) Unreliable matching pairs, which mainly come from the inherent data noise introduced by the significant intra-class variation and small inter-class variation. As illustrated in Figure 1(c), the selection of cross-modal matches based solely on the similarity may be improper. This arises from the fact that certain negative samples can be erroneously identified as ground truth due to their similarity with the target image. These issues collectively diminish the accuracy of cross-modal matches from different perspectives.

Motivated by the above observation, we propose a novel framework termed *Adaptive Uncertainty-based Learning (AUL)* for text-based person retrieval from the uncertainty perspective. In detail, as is depicted in Figure 2, it consists of three key components: 1) Uncertainty-aware Matching Filtration (UMF), which initially employs the Subjective Logic theory (Jøsang 2016) to model the uncertainty for measuring the degree of matching ambiguity, termed as matching uncertainty. Subsequently, it utilizes this uncertainty to adaptively assign weights to each training pair, which aims to prevent the impact of matching ambiguity and select high-confidence cross-modal matches during the model learning process. 2) Uncertainty-based Alignment Refinement (UAR), which not only explores the one-to-many correspondence through the construction of uncertainty representations but also engages in progressive learning to properly integrate coarse- and fine-grained alignments. Note that the concept of one-to-many correspondence can be likened to coarse-grained alignments. This module adeptly addresses the deficiency in one-to-many correspondence and guides the model in gradually acquiring more comprehensive alignments, through an easy-to-hard learning approach. 3) Cross-modal Masked Modeling (CMM) that designs masked signal modeling with cross-modal interaction, which effectively mines fine-grained relations between image and text. We evaluate our method on three widely used benchmarks for text-based person retrieval, *i.e.*, CUHK-PEDES, ICFG-PEDES, and RSTPReid. The experimental results suggest that our AUL method significantly outperforms recent state-of-the-art methods in supervised, weakly supervised, and domain generalization settings.

Our primary contributions can be summarized as follows:

- By carefully considering the matching uncertainty, we design an Uncertainty-aware Matching Filtration strategy, which leverages Subjective Logic to adaptively select high-confidence cross-modal matches and mitigate the disturbance of unreliable matching pairs for training.
- We propose an Uncertainty-based Alignment Refinement module, which not only simulates coarse-grained alignments by constructing uncertainty representations but also progressively organizes multi-grained alignments.
- We deploy a Cross-modal Masked Modeling module to reconstruct both image and text modality signals through comprehensive cross-modal interaction, which explores further correspondences between two modalities.

Related Work

Text-based Person Retrieval. The objective of text-based person retrieval is to accurately identify the target pedestrian based on the provided text. (Li et al. 2017) first introduced this task and released the pioneering dataset, CUHK-PEDES. Most subsequent methods (Niu et al. 2020; Niu, Huang, and Wang 2020; Zheng et al. 2020; Jing et al. 2020; Ding et al. 2021; Suo et al. 2022; Farooq et al. 2022; Wang et al. 2020; Aggarwal, Babu, and Chakraborty 2020) largely relied on the attention module or supplementary information to achieve effective cross-modal alignments. For example, (Wu et al. 2021) employed color reasoning to obtain informative semantics. (Farooq et al. 2022; Li et al. 2023b) designed a unified multi-layer network to dynamically extract the global- and local-level semantics from image and text modalities. Moreover, recent approaches (Li et al. 2023c; Jiang and Ye 2023) gradually utilized visual-language pre-training models or pre-training based on external knowledge for enhanced alignment capabilities. (Yan et al. 2022a) effectively harnessed the advantage of CLIP (Radford et al. 2021) for text-based person retrieval, while (Jiang and Ye 2023) improved the retrieval performance by pre-training on their constructed dataset. However, they overlook the disturbance of matching uncertainty arising from unreliable matching pairs, which forms the motivation of our work.

Uncertainty-based Learning. To address the challenge of quantifying prediction confidence, uncertainty-based learning has emerged as a promising approach. (Kendall and Gal 2017) have categorized uncertainty into two distinct types: epistemic uncertainty and aleatoric uncertainty. To tackle epistemic uncertainty, some studies have utilized the Bayesian network (Gal and Ghahramani 2016) and Subjective Logic (Jøsang 2016) with Dempster-Shafer theory of evidence (Yager and Liu 2008), aiming to learn the distribution of weights rather than obtain specific weights directly. In terms of aleatoric uncertainty, prior studies have explored it in various domains, such as image retrieval (Warburg et al. 2021), and segmentation(Zheng and Yang 2021). Different from these, we proposed an Adaptive Uncertainty-based Learning framework to surpass the disturbance of matching uncertainty and mine one-to-many correspondence by constructing uncertainty-based representations.

Masked Signal Modeling. (He et al. 2022; Vaswani et al. 2017; Jiang et al. 2023) have suggested that Masked Signal Modeling is a commonly used component in various visual and language tasks, comprising Masked Language Modeling (MLM) and Masked Image Modeling (MIM). For example, (Vaswani et al. 2017) verified the generalizability of MLM across a wide spectrum of natural language processing tasks. (He et al. 2022) predicted the masked pixels to refine visual representations. As their remarkable performance, MLM and MIM also take a significant part in our task. (Jiang and Ye 2023) introduced a method to predict masked textual tokens according to the unmasked textual

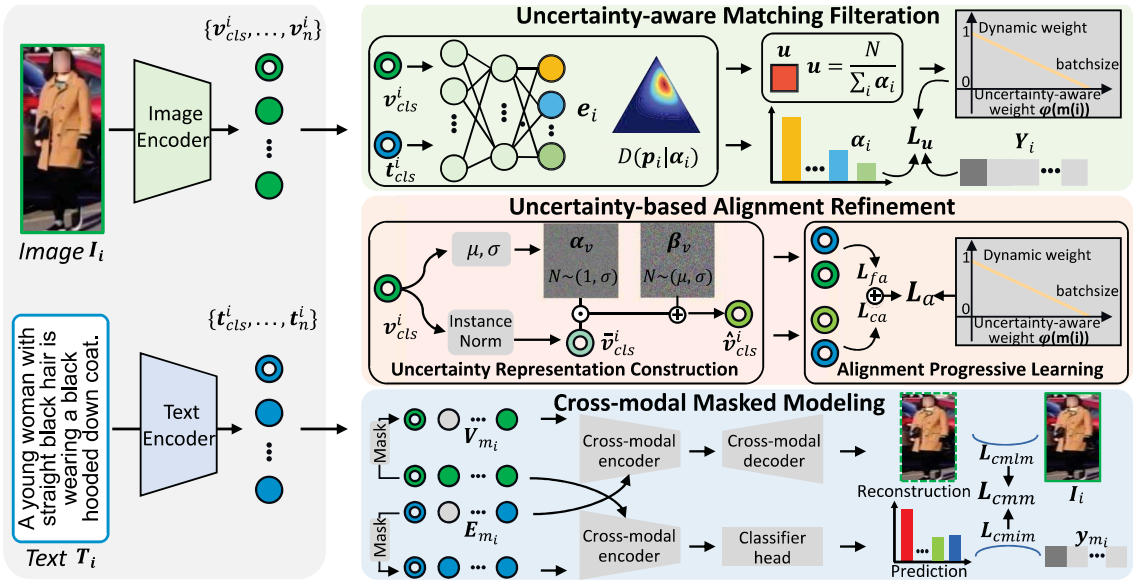

Figure 2. The overall framework of our proposed AUL method. It consists of three key components: 1) Uncertainty-aware Matching Filteration (UMF); 2) Uncertainty-based Alignment Refinement (UAR); and 3) Cross-modal Masked Modeling.

and visual tokens. In this paper, we predict masked textual and visual tokens by considering both modality semantics.

Our AUL Method

Preliminary

The objective of the text-based person retrieval task is to discern and retrieve the most similar person image from a candidate gallery, guided by a provided textual query. To obtain the correct pedestrian, our proposed framework focuses on facilitating accurate alignments by learning the similarity existing between the textual description and the corresponding person image. Formally, we define $\{\mathbf{I}_i, \mathbf{T}_i\}$ as an image-text pair within the training dataset. Each pair consists of a person image $\mathbf{I}_i$ and its corresponding textual description $\mathbf{T}_i$. We first input the image $\mathbf{I}_i$ into the image encoder, yielding a sequence of visual features $\{\mathbf{v}_i^{cls}, \mathbf{v}_i^1, \cdots, \mathbf{v}_i^n\}$, where $\mathbf{v}_i^{cls}$ served as the global visual feature, and $\{\mathbf{v}_i^1, \cdots, \mathbf{v}_i^n\}$ denote visual patch features. Moreover, we leverage the text encoder to obtain a sequence of textual representations $\{\mathbf{t}_i^{cls}, \mathbf{t}_i^1, \cdots, \mathbf{t}_i^n\}$, where $\mathbf{t}_i^{cls}$ and $\{\mathbf{t}_i^1, \cdots, \mathbf{t}_i^n\}$ represent the global textual feature and token features.

Uncertainty-aware Matching Filtration

Background of Subjective Logic. Subjective Logic (SL) offers a formalized representation of Dempster-Shafer (Yager and Liu 2008) theory's principle of uncertainty assignments within a discernment frame, modeled as a Dirichlet Distribution. Consequently, it provides the means to employ the principles of SL theory for quantifying the uncertainty, within a rigorously established theoretical framework. Specifically, we first obtain the evidence vector $\mathbf{e}_i$ predicted for the i-th singleton. Then we model the uncertainty u and belief mass $\mathbf{p} = \{p_k\}_{k=1}^N$ of each singleton, which can be formulated as follows:

$$p_k = \frac{e_k}{S}, \quad u = \frac{N}{S}, \tag{1}$$

where $S = \sum_{k=1}^{N}(e_k + 1)$ can be considered as the intensity of Dirichlet distribution, and the belief probability p_k corresponds to the parameters of the corresponding Dirichlet distribution $\boldsymbol{\alpha} = \{e_k + 1\}_{k=1}^N$. Note that the uncertainty u exhibits an inverse relationship with the total evidence. Finally, the Dirichlet distribution characterized by $\boldsymbol{\alpha}$ can be defined as:

$$D(\mathbf{p}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{N} p_j^{\alpha_j - 1} & \text{for } \mathbf{p} \in \mathcal{S}_N, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $B(\boldsymbol{\alpha})$ represents the N-dimensional beta function, and $\mathcal{S}_N$ is the N-dimensional unit simplex.

Uncertainty-aware Learning. In order to effectively mitigate the impact of uncertainty arising from unreliable matching pairs, the need to model matching uncertainty is evident. While the Subjective Logic (SL) theory has showcased remarkable advancements in uncertainty modeling, it is unsuitable to consider the direct application of the SL to text-based person retrieval. To expand the SL to this specific task, the initial step is to denote the prediction of cross-modal match evidence $e_{ij} = exp^{f(Sim(\mathbf{t}_i^{cls}, \mathbf{v}_i^{cls}))}$ between i-th text and j-th image, where $Sim(\cdot)$ and f denote the calculation of cosine similarity and the ReLU function. The evidence $\mathbf{e}_i$ of the total matches of i-th text can be expressed as $\mathbf{e}_i = \{e_{ij}\}_{j=1}^N$.

Following the Subjective Logic mentioned in the previous section, we obtain $\boldsymbol{\alpha}_i$ and model the matching uncertainty $\mathbf{u}$ as follows:

$$\boldsymbol{\alpha}_i = \mathbf{e}_i + 1, \quad \mathbf{u} = \frac{N}{\mathbf{S}}, \tag{3}$$

where $\mathbf{S} = \sum_{i=1}^{N}(\boldsymbol{\alpha}_i)$ can be viewed as the intensity of Dirichlet distribution. Based on the obtained matching uncertainty, we perform Uncertainty-aware Learning to adaptively filter unreliable matching pairs and select high-confidence cross-modal matches. Specifically, we design the cross-entropy loss $\mathcal{L}_u$ with uncertainty-aware dynamic weight function $\varphi(m(i))$ to assign larger weights to cross-modal matches with lower matching uncertainty and smaller weights to ones with higher matching uncertainty in the optimization process, thus reducing the negative impact caused by unreliable matching pairs. The loss function $\mathcal{L}_u$ can be represented as follows:

$$\mathcal{L}_u = \lambda \sum_{i=1}^{N} \varphi(m(i)) \mathbf{Y}_i \Big(\log(S_i) - \log(\boldsymbol{\alpha}_i) \Big), \tag{4}$$

where λ is a hyper-parameter and $\mathbf{Y}_i$ is a one-hot label for i-th sample, and $\varphi(m(i)) = \dfrac{m(i)}{N} \in (0, 1]$, $m(i)$ indicates the ordinal number of i-th cross-modal match obtained by sorting the matching uncertainty $\mathbf{u}$ in descending order.

Uncertainty-based Alignment Refinement

Due to the absence of one-to-many correspondence between vision and language, existing methods mainly focus on exploring one-sided cross-modal alignment, *i.e.*, one-to-one correspondence, leading to a degenerated retrieval performance. To address this limitation, we propose an Uncertainty-based Alignment Refinement (UAR) module that simulates coarse-grained alignments and employs progressive learning to collaboratively refine coarse- and fine-grained alignments in an easy-to-hard manner.

Uncertainty Representation Construction. Given the global representations $(\mathbf{v}^{cls}, \mathbf{t}^{cls})$ of N image-text pairs, we need to explicitly construct visual representations with uncertainty first, which is achieved by appending the Gaussian Noise of the original feature distribution. The mean μ and standard deviation σ of the Gaussian Noise are derived from the original features $\mathbf{v}^{cls}$. We then construct visual representations with uncertainty $\hat{\mathbf{v}}^{cls}$ by adding the generated Gaussian Noise to the whitened features $\bar{\mathbf{v}}^{cls}$, which can be formulated as follows:

$$\hat{\mathbf{v}}^{cls} = \boldsymbol{\alpha}_v \cdot \bar{\mathbf{v}}^{cls} + \boldsymbol{\beta}_v, \tag{5}$$

where $\boldsymbol{\alpha}_v$ and $\boldsymbol{\beta}_v$ are the uncertainty vectors introducing noise, $\boldsymbol{\alpha}_v \sim N(1, \sigma), \boldsymbol{\beta}_v \sim N(\mu, \sigma)$, and $\bar{\mathbf{v}}^{cls}$ is whitened feature $\bar{\mathbf{v}}^{cls} = \frac{\mathbf{v}^{cls} - \mu}{\sigma}$.

Alignment Progressive Learning. Based on the obtained visual representations with uncertainty $\hat{\mathbf{v}}_i^{cls}$ and textual representation $\mathbf{t}_i^{cls}$, we adopt InfoNCE loss $\mathcal{L}_{info}$ (Lee, Kim, and Han 2021; Yang et al. 2023) to perform coarse-grained alignments and further explore the one-to-many correspondence. The loss for coarse-grained alignments can be defined as follows:

$$\mathcal{L}_{ca} = \frac{\mathcal{L}_{info}\left(\hat{\mathbf{v}}_i^{cls},\ \mathbf{t}_i^{cls}\right)}{2\sigma^2} + \frac{1}{2}\log\sigma^2, \tag{6}$$

In terms of the fine-grained alignments, *i.e.*, one-to-one correspondence, we design a pair-wise loss function $\mathcal{L}_{fa}$ to alleviate the adverse effect (Zhou et al. 2023) of dense sampling mechanism. The pair-wise loss function using only one negative sample $\mathbf{t}_{neg}^{cls}$ can be written as:

$$\mathcal{L}_{fa} = -\log \frac{\psi\left(\mathbf{v}_i^{cls}, \mathbf{t}_i^{cls}\right)}{\psi\left(\mathbf{v}_i^{cls}, \mathbf{t}_{neg}^{cls}\right) + \psi\left(\mathbf{v}_i^{cls}, \mathbf{t}_i^{cls}\right)}, \tag{7}$$

Intuitively, conducting fine-grained alignments is notably more challenging than coarse-grained alignments. Therefore, our strategy involves assigning higher weights to coarse-grained alignments and lower weights to fine-grained alignments at the outset, gradually reversing this allocation during the training process. We propose the Alignment Progressive Learning (APL) to incorporate dynamic weights into the loss function, allowing a gradual focus on multi-grained alignment in an "easy-to-hard" manner while optimizing the following objective $\mathcal{L}_a$:

$$\mathcal{L}_a = \sum_{i=1}^{N} \varphi(m(i))(\gamma \mathcal{L}_{ca} + (1-\gamma)\,\mathcal{L}_{fa}), \tag{8}$$

where $\gamma = \exp(-\gamma_0 \cdot \frac{epoch}{total\_epoch})$, and γ_0 is initial weight.

Cross-modal Masked Modeling

To enhance the interaction between image and text, we design the Cross-modal Masked Modeling (CMM) to reconstruct the inherent signals of one modality using a masked input, which is conditioned on the unmasked inputs of both the image and text modalities. This CMM can be further divided into two components: Cross-modal Masked Image Modeling (CMIM) and Cross-modal Masked Language Modeling (CMLM).

Taking CMIM for an example, following MAE (He et al. 2022), we obtain the representation $\mathbf{V}_{m_i} = \{\mathbf{v}_i^j\}_{j=1}^{n_u}$ of a masked image, n_u denotes the number of unmasked tokens. Then we utilize the cross-modal encoder f_e including a multi-head cross attention layer and 3-layer transformer blocks, to obtain the prediction of all original tokens according to the representation of masked image $\mathbf{V}_{m_i}$ and original textual representation $\mathbf{E}_i = \{\mathbf{t}_i^j\}_{j=1}^{n}$. Finally, The prediction is mapped back to the RGB image space by an image cross-modal decoder f_d, of which the structure is the same as the encoder and followed by a linear layer. The total procedure of CMIM is represented as:

$$\mathcal{L}_{cmim} = \frac{1}{\Omega(\mathbf{I}_i)} \|\mathbf{I}_i - f_d(f_e(\mathbf{V}_{m_i}, \mathbf{E}_i))\|_1, \tag{9}$$

where $\Omega(\cdot)$ is the number of pixels, and the loss function $\mathcal{L}_{cmim}$ is based on the l_1 loss.

Similar to CMIM, given the representation $\mathbf{E}_{m_i}$ of a masked text and original visual representation $\mathbf{V}_i$, we utilize the cross-entropy loss function $\mathcal{H}$ to measure the distance between predictions and masked textual tokens $\mathbf{E}_{m_i}$, *i.e.*, performing Cross-modal Masked Language Modeling. Therefore, the objective of CMM can be calculated as:

$$\mathcal{L}_{cmm} = \mathcal{L}_{cmim} + \mathcal{H}(\mathbf{y}_{m_i}, f_{td}(f_{te}(\mathbf{V}_i, \mathbf{E}_{m_i})), \tag{10}$$

where $\mathbf{Y}_{m_i}$ is the one-hot label of i-th masked token, f_{te} is the same as the cross-modal encoder of CMIM, and f_{td} is a Classifier head. By minimizing $\mathcal{L}_{cmm}$, the model is compelled to perform the reconstruction of original signals through cross-modal interaction. This process efficiently facilitates the exploration of deeper relations existing between the image and text modalities.

Finally, the overall loss $\mathcal{L}_{total}$ for training is denoted as:

$$\mathcal{L}_{total} = \mathcal{L}_u + \mathcal{L}_a + \mathcal{L}_{cmim}. \tag{11}$$

Experiments

Experimental Setup

Datasets. We evaluate our model on three benchmark datasets, including: *1) CUHK-PEDES* (Li et al. 2017) encompasses a total of 40,206 images, capturing 13,003 distinct identities and accompanied by 68,120 textual descriptions. *2) ICFG-PEDES* (Ding et al. 2021) is a large-scale person dataset, comprising a substantial collection of 54,522 images, of which the training set contains 34,674 image-text pairs, while the test set includes 19,848 image-text pairs. *3) RSTPReid* (Zhu et al. 2021) comprises a total of 20,505 images spanning 4,101 distinct identities, all captured by 15 different cameras. Additionally, every image is enriched with two textual descriptions.

Evaluation. Following previous approaches (Jiang and Ye 2023; Yang et al. 2023), we adopt Rank@K (R@K) as the standard evaluation for all datasets, where is the percentage of retrieving at least a singular corresponding target image from top-k candidate images.

Implementation Details. We implement our model with PyTorch. For fairness, we follow the VLP-based method (Yang et al. 2023) and employ the same visual and textual encoder, *i.e.*, Swin Transformer (Liu et al. 2021) and Bert (Vaswani et al. 2017). The total training procedure can be divided into two stages. We first train the whole model on the dataset proposed by (Yang et al. 2023). Then we resize the image to 384 $\times$ 128 and set the length for each textual token sequence to 56. Initialed by parameters of the first stage, we trained our AUL model with PyTorch for 35 epochs using the Adam optimizer (Kingma and Ba 2015) with a learning rate initialed by $5e$-5 and decayed to $5e$-6 following a linear learning rate decay. The batch size is set as 128. Finally, λ and γ_0 are set to 0.8 and 1.0 for all experiments.

Overall Comparsion Results

We compare our proposed method CMAP with recent state-of-the-art methods, including: *(1) Traditional pre-training methods* that improve the accuracy of cross-modal matches by attention mechanisms and additional informative cues, such as LGUR (Shao et al. 2022), LBUL (Wang et al. 2022b); Several methods DSSL (Zhu et al. 2021), AXM-Net (Farooq et al. 2022), ISANet (Yan et al. 2022b), CAIBC (Wang et al. 2022a), RKT (Wu et al. 2023) and SRCF (Suo et al. 2022) propose some simple strategies to achieve distinct semantics for proper alignments. *(2) Vision-language pre-training methods* that leverage prior knowledge from extra large image-text corpora, including the CFine (Yan et al. 2022a), TP-TPS (Wang et al. 2023), IRRA (Jiang and Ye 2023), RaSa (Bai et al. 2023), and APTM (Yang et al. 2023).

	Method	R@1	R@5	R@10
w/o VLP	DSSL (MM'21)	59.98	80.41	87.56
	SSAN (arXiv'21)	61.37	80.15	86.73
	AXM-Net (AAAI'22)	61.90	79.40	85.75
	CAIBC (MM'22)	64.43	82.87	88.37
	LBUL (MM'22)	64.04	82.66	87.22
	LGUR (MM'22)	64.21	81.94	87.93
	C_2A_2 (MM'22)	64.82	83.54	89.77
	ISANet (arXiv'22)	63.92	82.15	87.69
	SRCF (ECCV'22)	64.04	82.99	88.81
	RKT (TMM'23)	61.48	80.74	87.28
	ASAMN (TIP'23)	65.66	84.53	90.21
w/ VLP	IVT (ECCVW'22)	65.59	83.11	89.21
	CFine (arXiv'22)	69.57	85.93	91.15
	TP-TPS (arXiv'23)	70.16	86.10	90.98
	IRRA (CVPR'23)	73.38	89.93	93.71
	RaSa (IJCAI'23)	76.51	90.29	94.25
	APTM (MM'23)	76.17	89.47	93.57
	AUL (Ours)	**77.23**	**90.43**	**94.41**

Table 1. Comparisons on CUHK-PEDES.

	Method	R@1	R@5	R@10
w/o	DSSL (MM'21)	39.05	62.60	73.95
	SSAN (arXiv'21)	43.50	67.80	77.15
	LBUL (MM'22)	45.55	68.20	77.85
	C_2A_2 (MM'22)	51.55	76.75	85.15
w/ VLP	IVT (ECCVW'22)	46.70	70.00	78.80
	CFine (arXiv'22)	50.55	72.50	81.60
	TP-TPS (arXiv'23)	50.65	72.45	81.20
	IRRA (CVPR'23)	60.20	81.30	88.20
	RaSa (IJCAI'23)	66.90	86.50	91.35
	APTM (MM'23)	66.45	85.60	90.60
	AUL (Ours)	**71.65**	**87.55**	**92.05**

Table 2. Comparisons with recent methods on RSTPReid.

Comparison on Supervised Setting. According to the comparison on three datasets reported in Table 1, 2, and 3, we can find that: (1) Specifically, our AUL model achieves a remarkable 71.65% R@1 on RSTPReid, outperforming RaSa and APTM by 4.75% and 5.20%. These results suggest that our AUL method excels in relieving the adverse impact of severe matching ambiguity and mining more fine-grained correspondence between vision and language. (2) Furthermore, compared to current methods, we achieve significant improvements on all three datasets, which indicates that our AUL model effectively refines cross-modal alignments by simulating coarse-grained alignments and progressively collaborates multi-grained alignments.

Comparison on Weakly Supervised and Domain Generalization Settings. Additionally, we also evaluate our AUL model in weakly supervised and domain generalization settings. From the Table 5 and 6, we can observe that: (1) Our AUL model exhibits a substantial performance gain over the current state-of-the-art APTM method, particularly in terms of R@1. This improvement can be attributed to the fact that in scenarios where only pairwise relationships are

	Method	R@1	R@5	R@10
w/o VLP	MIA (TIP'20)	46.49	67.14	75.18
	SSAN (arXiv'21)	54.23	72.63	79.53
	LGUR (MM'22)	57.42	74.97	81.45
	ISANet (arXiv'22)	57.73	75.42	81.72
	SRCF (ECCV'22)	57.18	75.01	81.49
	ASAMN (TIP'23)	57.09	76.33	82.84
w/ VLP	IVT (ECCVW'22)	56.04	73.60	80.22
	CFine (arXiv'22)	60.83	76.55	82.42
	TP-TPS (arXiv'23)	60.64	75.97	81.76
	IRRA (CVPR'23)	63.46	80.25	85.82
	RaSa (IJCAI'23)	65.28	80.40	85.12
	APTM (MM'23)	68.22	82.87	87.50
	AUL (Ours)	**69.16**	**83.32**	**88.37**

Table 3. Comparisons on ICFG-PEDES.

No.	Components				RSTPReid		
	UMF	UAR		CMM	R@1	R@5	R@10
		$\mathcal{L}_{fa}$	$\mathcal{L}_{ca}$				
0	-	-	-	-	68.15	85.10	89.20
1	-	-	-	✔	69.25	85.65	90.20
2	-	✔	✔	-	69.15	86.10	90.35
3	✔	-	-	-	69.55	85.40	89.50
4	-	✔	-	✔	70.85	85.95	90.55
5	-	✔	✔	✔	70.95	87.10	91.25
6	✔	-	-	✔	71.35	86.85	90.90
7	✔	✔	✔	✔	**71.65**	**87.55**	**92.05**

Table 4. Ablation studies with respect to model components on RSTPReid.

Methods	R@1	R@5	R@10
CMPM+MMT (ICCV'21)	50.51	70.23	78.98
CMPM+SpCL (ICCV'21)	51.13	71.54	80.03
CMMT (ICCV'21)	57.10	78.14	85.23
CAIBC (MM'22)	58.64	79.02	85.93
IRRA (CVPR'23)	70.94	88.39	93.06
APTM (MM'23)	74.57	88.95	93.18
AUL (Ours)	**75.86**	**90.11**	**94.02**

Table 5. Comparisons with state-of-the-arts (weakly supervised) on CUHK-PEDES.

	Methods	R@1	R@5	R@10
C → I	SSAN (arXiv'21)	29.24	49.00	58.53
	LGUR (MM'22)	34.25	52.58	60.85
	C_2A_2 (MM'22)	27.61	47.48	57.03
	ASAMN (TIP'23)	30.22	50.51	59.59
	IRRA (CVPR'23)	41.89	61.56	69.04
	APTM (MM'23)	46.20	65.13	72.59
	AUL (Ours)	**49.29**	**67.46**	**74.42**
I → C	SSAN (arXiv'21)	21.07	38.94	48.54
	LGUR (MM'22)	25.44	44.48	54.39
	C_2A_2 (MM'22)	16.48	34.03	43.88
	ASAMN (TIP'23)	17.99	35.30	44.75
	IRRA (CVPR'23)	31.04	52.18	63.53
	APTM (MM'23)	48.67	68.75	77.06
	AUL (Ours)	**56.79**	**76.14**	**83.14**

Table 6. Comparisons with state-of-the-arts (domain generalization). Here "C" denotes CUHK-PEDES, while "I" represents ICFG-PEDES.

available and identity information is absent, the impact of significant intra-class variations becomes more pronounced, resulting in compromised retrieval performance. Hence, we proposed UMF to alleviate the influence of matching ambiguity. (2) Moreover, our AUL model achieves an improvement of 8.12% in R@1 and 7.39% in R@5 over the recent APTM method, in terms of the I → C. Such results demonstrate that our AUL model can effectively quantify the uncertainty inherent in cross-modal matching ambiguity and filter out high-confidence alignments.

Further Analysis

Ablation Study. As illustrated in Table 4, we list the following conclusions: (1) The comparison with No.0 and No.3 reveals that our proposed UMF significantly enhances retrieval performance. Such demonstrates again that introducing the SL theory to model the uncertainty of cross-modal matching ambiguity is effective for filtering the high-confidence alignment, which makes our model dedicated to reliable retrieval results. (2) The model performance of No.5 is better than the result of No.1, especially in terms of R@5 and R@10. It indicates that UAR can effectively explore one-to-many correspondence through the application of Gaussian Noise-based uncertainty representation. Additionally, the progressive learning approach employed by UAR appropriately collaborates both coarse- and fine-grained alignments. (3) From the comparison of No.6 and No.3, we speculate that Adding the CMM has a greater impact on the retrieval performance. One probable reason is that performing MLM and MIM with further cross-modal interaction yields additional advantages in terms of fine-grained and relevant relation mining between vision and language.

Analysis on Choice of CMLM and CMIM. We further explore the importance of CMLM and CMIM respectively. As shown in Figure 4, we can observe that: (1) The ablated model w/ $\mathcal{L}_{cmlm}$ performs better than baseline. We consider that the superior performance is due to the full interaction between image and text, which is more helpful to bridge the significant modality gap between vision and language. (2) Furthermore, applying the loss $\mathcal{L}_{cmlm}$ of CMLM solely is not effective as the combination of CMIM and CMLM, *i.e.*, w/ CMM. It suggests that using both masked textual and visual tokens as anchors for mining comprehensive cross-modal relations is indispensable.

Analysis on Alignment Progressive Learning in UAR. Here, we study the advancement of our proposed Alignment Progressive Learning (APL), which aims to comprehensively explore one-to-one and one-to-many correspondence. By observing Figure 3, we can find that: (1) Introducing the dynamic weight γ performs better than the ablated model of Avg. We speculate the reason is that the utilization of progressive learning plays a significant role in learning comprehensive multi-grained alignments. (2) The proposed APL effectively allocates higher weights to coarse-grained alignments initially at the out, gradually shifting to allocate

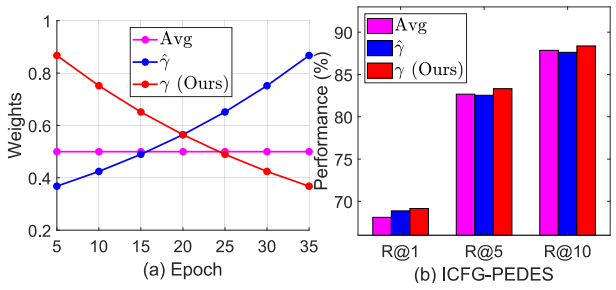

Figure 3. Analysis with respect to the Alignment Progressive Learning in UAR. γ and $\hat{\gamma}$ denote our proposed progressive learning manner and the reverse manner respectively, and Avg is the average weight assignment.

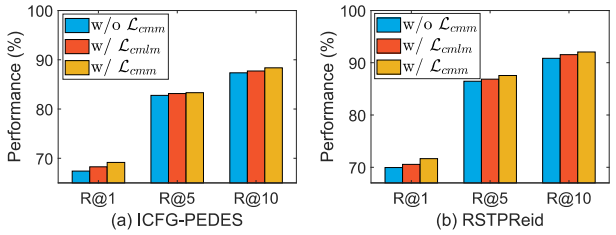

Figure 4. Effect of the components in Cross-modal Masked Modeling on ICFG-PEDES and RSTPReid.

higher weights to fine-grained alignments. (3) Moreover, we further explore the effectiveness of learning multi-grained alignments in an easy-to-hard manner. In particular, we compare the performance of leveraging γ and $\hat{\gamma}$. Obviously, the former exhibits better suitability for proper alignment incorporation and retrieval accuracy. This finding supports our intuition that guiding the model to progressively learn appropriate multi-grained alignments in an easy-to-hard manner is more reasonable than others.

Analysis on Matching Uncertainty-aware Dynamic Weight Assignment. To further validate the existence of matching ambiguity and the significance of our proposed UMF, we conducted an in-depth investigation into the relationship between various uncertainty-aware weight allocations and the overall performance. The observations drawn from Figure 5 are outlined as follows: (1) The distribution analysis clearly reveals the existence of unreliable matching pairs, characterized by pronounced matching uncertainty. This uncertainty arises from significant intra-class variations and limited inter-class variations, impeding the enhancement of retrieval performance. (2) In order to underscore the efficacy of matching uncertainty-aware dynamic weight assignments, we compare the performance of diverse weight assignments. Setting high uncertainty cross-modal matches as $1(\cdot)$ yields the poorest performance, which reflects the rationality of our motivation that our model suffers from severe matching ambiguity.

Qualitative Analysis. As shown in Figure 6, we present a qualitative analysis that compares the top-6 retrieved results of our AUL method with the recent APTM method (Yang et al. 2023). According to the visualization results,

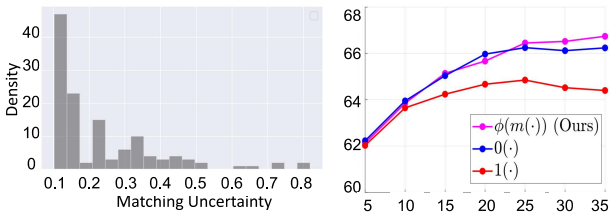

Figure 5. The effectiveness of matching uncertainty-aware weight assignment on ICFG-PEDES. $0(\cdot)$ and $1(\cdot)$ denote the weight setting of samples to 0 and 1 respectively when the uncertainty is greater than 0.5.

Figure 6. Qualitative results of APTM and our proposed AUL on ICFG-PEDES and CUHK-PEDES.

our AUL reflects the superiority in retrieval accuracy over the APTM method. Specifically, our method AUL can satisfy both the fine-grained and coarse-grained retrieval requirement, such as *"long sleeve"* or *"tall man"*), due to the fact that our proposed UAR acquire multi-grained semantics progressively and comprehensively. Moreover, the AUL also models the matching uncertainty to quantify the ambiguity caused by large intra-class variation and minimal inter-class variation, which mitigates the disturbance of unreliable matching pairs, thus improving the performance.

Conclusion

In this paper, we proposed a novel Adaptive Uncertainty-based Learning (AUL) method for text-based person retrieval from an uncertainty perspective. We proposed the Uncertainty-aware Matching Filtration (UMF) to quantify and prevent the influence of ambiguity caused by unreliable matching pairs. Moreover, we design Uncertainty-based Alignment Refinement (UAR) and Cross-modal Masked Modeling (CMM) to enhance alignment learning and focus on proper cross-modal relations. Extensive experiments conducted on three benchmarks demonstrate the superiority of our proposed AUL method. In the future, we will explore other strategies to enhance the retrieval performance.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grants (No. 62222203, 62072080 and U20B2063) and the New Cornerstone Science Foundation through the XPLORER PRIZE, and the Science and Technology Innovation Committee of Shenzhen Municipality Foundation (No. JCYJ20210324132203007).

References

Aggarwal, S.; Babu, R. V.; and Chakraborty, A. 2020. Text-based Person Search via Attribute-aided Matching. In *IEEE Winter Conference on Applications of Computer Vision*, 2606–2614.

Bai, Y.; Cao, M.; Gao, D.; Cao, Z.; Chen, C.; Fan, Z.; Nie, L.; and Zhang, M. 2023. RaSa: Relation and Sensitivity Aware Representation Learning for Text-based Person Search. *CoRR*, abs/2305.13653.

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification. *CoRR*, abs/2107.12666.

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2022. AXM-Net: Implicit Cross-Modal Feature Alignment for Person Re-identification. In *AAAI*, 4477–4485.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the International Conference on Machine Learning*, volume 48, 1050–1059.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. B. 2022. Masked Autoencoders Are Scalable Vision Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15979–15988.

Hou, R.; Chang, H.; Ma, B.; Huang, R.; and Shan, S. 2021. BiCnet-TKS: Learning Efficient Spatial-Temporal Representation for Video Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014–2023.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. *CoRR*, abs/2303.12501.

Jiang, X.; Xu, X.; Zhang, J.; Shen, F.; Cao, Z.; and Shen, H. T. 2022. Semi-supervised video paragraph grounding with contrastive encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2466–2475.

Jiang, X.; Zhou, Z.; Xu, X.; Yang, Y.; Wang, G.; and Shen, H. T. 2023. Faster Video Moment Retrieval with Point-Level Supervision. *arXiv preprint arXiv:2305.14017*.

Jing, Y.; Si, C.; Wang, J.; Wang, W.; Wang, L.; and Tan, T. 2020. Pose-Guided Multi-Granularity Attention Network for Text-Based Person Search. In *AAAI*, 11189–11196.

Jøsang, A. 2016. *Subjective Logic - A Formalism for Reasoning Under Uncertainty*. Springer.

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, 5574–5584.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Lee, S.; Kim, D.; and Han, B. 2021. CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback. In *CVPR*, 802–812.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person Search with Natural Language Description. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5187–5196.

Li, S.; Xu, X.; Jiang, X.; Shen, F.; Liu, X.; and Shen, H. T. 2023a. Multi-Grained Attention Network with Mutual Exclusion for Composed Query-Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, S.; Xu, X.; Shen, F.; and Yang, Y. 2023b. Multi-granularity Separation Network for Text-Based Person Retrieval with Bidirectional Refinement Regularization. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 307–315.

Li, S.; Xu, X.; Yang, Y.; Shen, F.; Mo, Y.; Li, Y.; and Shen, H. T. 2023c. DCEL: Deep Cross-modal Evidential Learning for Text-Based Person Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6292–6300.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *IEEE/CVF International Conference on Computer Vision*, 9992–10002.

Niu, K.; Huang, Y.; Ouyang, W.; and Wang, L. 2020. Improving Description-Based Person Re-Identification by Multi-Granularity Image-Text Alignments. *IEEE Trans. Image Process.*, 29: 5542–5556.

Niu, K.; Huang, Y.; and Wang, L. 2020. Textual Dependency Embedding for Person Search by Language. In Chen, C. W.; Cucchiara, R.; Hua, X.; Qi, G.; Ricci, E.; Zhang, Z.; and Zimmermann, R., eds., *ACM International Conference on Multimedia*, 4032–4040.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning Granularity-Unified Representations for Text-to-Image Person Re-identification. In *The ACM International Conference on Multimedia*, 5566–5574.

Specker, A.; Cormier, M.; and Beyerer, J. 2023. UPAR: Unified Pedestrian Attribute Recognition and Person Retrieval. In *WACV*, 981–990.

Suo, W.; Sun, M.; Niu, K.; Gao, Y.; Wang, P.; Zhang, Y.; and Wu, Q. 2022. A Simple and Robust Correlation Filtering Method for Text-Based Person Search. In *ECCV*, 726–742.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, G.; Yu, F.; Li, J.; Jia, Q.; and Ding, S. 2023. Exploiting the Textual Potential from Vision-Language Pre-training for Text-based Person Search. *CoRR*, abs/2303.04497.

Wang, Z.; Fang, Z.; Wang, J.; and Yang, Y. 2020. ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII*, volume 12357, 402–420.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022a. CAIBC: Capturing All-round Information Beyond Color for Text-based Person Retrieval. In *ACM International Conference on Multimedia*, 5314–5322.

Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T.; and Li, Y. 2022b. Look Before You Leap: Improving Text-based Person Retrieval by Learning A Consistent Cross-modal Common Manifold. In *The ACM International Conference on Multimedia*, 1984–1992.

Warburg, F.; Jørgensen, M.; Civera, J.; and Hauberg, S. 2021. Bayesian Triplet Loss: Uncertainty Quantification in Image Retrieval. In *IEEE/CVF International Conference on Computer Vision*, 12138–12148.

Wu, Y.; Yan, Z.; Han, X.; Li, G.; Zou, C.; and Cui, S. 2021. LapsCore: Language-guided Person Search via Color Reasoning. In *IEEE/CVF International Conference on Computer Vision*, 1604–1613.

Wu, Z.; Ma, B.; Chang, H.; and Shan, S. 2023. Refined Knowledge Transfer for Language-Based Person Search. *IEEE Transactions on Multimedia*, 1–15.

Yager, R. R.; and Liu, L. 2008. *Classic Works of the Dempster-Shafer Theory of Belief Functions*, volume 219. Springer.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2022a. CLIP-Driven Fine-grained Text-Image Person Re-identification. *CoRR*, abs/2210.10276.

Yan, S.; Tang, H.; Zhang, L.; and Tang, J. 2022b. Image-Specific Information Suppression and Implicit Local Alignment for Text-based Person Search. *CoRR*, abs/2208.14365.

Yang, S.; Zhou, Y.; Wang, Y.; Wu, Y.; Zhu, L.; and Zheng, Z. 2023. Towards Unified Text-based Person Retrieval: A Large-scale Multi-Attribute and Language Search Benchmark. *CoRR*, abs/2306.02898.

Zheng, Z.; and Yang, Y. 2021. Rectifying Pseudo Label Learning via Uncertainty Estimation for Domain Adaptive Semantic Segmentation. *Int. J. Comput. Vis.*, 129(4): 1106–1120.

Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; and Shen, Y. 2020. Dual-path Convolutional Image-Text Embeddings with Instance Loss. *ACM Trans. Multim. Comput. Commun. Appl.*, 16: 51:1–51:23.

Zhou, X.; Zhong, Y.; Cheng, Z.; Liang, F.; and Ma, L. 2023. Adaptive Sparse Pairwise Loss for Object Re-Identification. *CoRR*, abs/2303.18247.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, 209–217.