From Data Deluge to Data Curation: A Filtering-WoRA Paradigm for Efficient Text-based Person Search

Jintao Sun
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
3120215524@bit.edu.cn

Hao Fei
School of Computing
National University of Singapore
Singapore, Singapore
haofei37@nus.edu.sg

Gangyi Ding
School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
dgy@bit.edu.cn

Zhedong Zheng*
Faculty of Science and Technology, and Institute of Collaborative Innovation
University of Macau
Macau, China
zhedongzheng@um.edu.mo

Abstract

In text-based person search endeavors, data generation has emerged as a prevailing practice, addressing concerns over privacy preservation and the arduous task of manual annotation. Although the number of synthesized data can be infinite in theory, the scientific conundrum persists that how much generated data optimally fuels subsequent model training. We observe that only a subset of the data in these constructed datasets plays a decisive role. Therefore, we introduce a new Filtering-WoRA paradigm, which contains a filtering algorithm to identify this crucial data subset and WoRA (Weighted Low-Rank Adaptation) learning strategy for light fine-tuning. The filtering algorithm is based on the cross-modality relevance to remove the lots of coarse matching synthesis pairs. As the number of data decreases, we do not need to fine-tune the entire model. Therefore, we propose a WoRA learning strategy to efficiently update a minimal portion of model parameters. WoRA streamlines the learning process, enabling heightened efficiency in extracting knowledge from fewer, yet potent, data instances. Extensive experimentation validates the efficacy of pretraining, where our model achieves advanced and efficient retrieval performance on challenging real-world benchmarks. Notably, on the CUHK-PEDES dataset, we have achieved a competitive mAP of 67.02% while reducing model training time by 19.82%.

CCS Concepts

• **Information systems** → **Retrieval efficiency**; **Retrieval effectiveness**; **Multimedia and multimodal retrieval**.

Keywords

Text-based Person Search, Data-centric Learning, Low-Rank Adaptation, Visual-language Pre-training
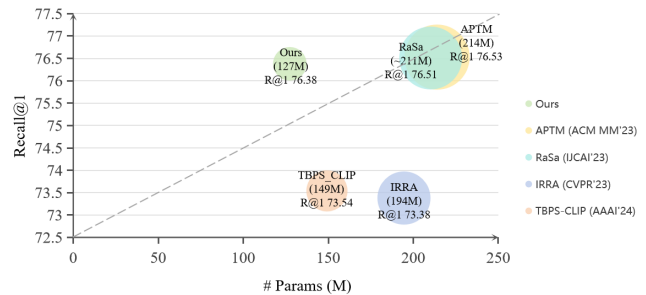
*Corresponding author.

Figure 1: Comparison between the proposed method and existing approaches in terms of Recall@1 and the parameter numbers. We observe that our method deploys fewer parameters while achieving a competitive Recall@1, *i.e.*, APTM [76], RaSa [4], IRRA [25] and TBPS-CLIP [5].

1 Introduction

Compared to traditional image-based person search [57, 63, 64, 66, 67, 85, 89], which seeks to retrieve target individuals from a vast array of footage or images across different locations and times, text-based person search locates interested individuals from a pool of candidates based on pedestrian descriptions [36]. Given that pedestrian image queries may not be available, text-guided person search emerges as an alternative method. The key lies in mining the fine-grained information from images and texts, blending the complexity of natural language processing with the subtle nuances of visual recognition, and establishing their correspondence. By leveraging the ability to understand complex human descriptions and accurately identify and retrieve images of individuals from a camera system, it can be applied to broad applications in public safety domains such as missing person searches [2, 59], and rescue operations [56]. As the sub-task of vision-language retrieval, text-based person search models usually require extensive data for training, where, however, the number of pedestrian data is limited.

Most datasets [13, 37, 76, 89] are constructed from three sources. (1) The first source is through sampling from camera footage, accompanied by manual annotations. However, constructing large-scale datasets is often infeasible due to privacy concerns and high costs. (2) The second source involves collecting images and short videos from the internet. Despite expanding dataset sizes, the noisy web text and the inconsistent quality of task images are generally suboptimal for fine-grained vision-language learning. (3) Therefore, most researchers [76] resort to leverage the generative models, *e.g.*,

GAN [87] and Diffusion [50, 55]. For instance, APTM [76] has introduced 1.51M image-text pairs generated by Stable Diffusion [50], showcasing the potential of training on a large synthesized dataset.

Despite significant progress in learning from large synthesized datasets, a fundamental challenge persists: **how can we efficiently extract knowledge when faced with an effectively infinite amount of data, considering the substantial computational costs incurred during training?** We observe two key points: (1) Previous works [76, 84, 87] indicate that performance gains diminish even with abundant generated data. This suggests that not all information in massive synthetic datasets is equally useful; instead, a well-chosen subset, or coreset, may be sufficient to capture the essential training information. (2) When learning from a coreset, updating the entire parameter space is unnecessary to achieve training precision. This insight opens the door to significantly reducing both training time and model complexity. Thus, our focus shifts to developing strategies that achieve high accuracy with reduced data and a compact parameter footprint.

To this end, we propose a new Filtering-WoRA paradigm, which contains a novel two-stage data filtration method aimed at identifying the coreset to enhance model performance and a WoRA (Weighted Low-Rank Adaptation) algorithm to optimize the pre-training and fine-tuning models, enabling training with fewer parameters while maintaining model performance and increasing computational speed. Specifically, our process begins with dataset purification, including the synthesized dataset for pre-training and the real-world dataset for fine-tuning. To filter out low-quality image-text pairs, *e.g.*, incomplete descriptive details or blurred image details, we leverage the off-the-shelf large cross-modality model to extract features from both images and texts within the dataset and then calculate the cosine similarity between projected image embeddings and projected text embeddings. This process yields a similarity score for each image-text pair, facilitating the selection of high-quality datasets based on our predetermined threshold. Subsequently, to reduce model parameters and enhance computational speed, we opt to freeze the weights from pre-training, indirectly training some dense layers in the neural network by optimizing rank decomposition matrices that change during the adaptation process. By decomposing pre-training and fine-tuning weights into magnitude and direction, our WoRA method introduces three new dimensions to facilitate the modification of the weight matrix and rank decomposition matrix. This approach allows for learning a minimal amount of parameters while simultaneously boosting model performance (see Figure 1). In summary, our contributions are:

- We propose Filtering-WoRA, a paradigm for efficient text-based person search that streamlines learning through focused data curation and targeted parameter updates. Our filtering algorithm assesses cross-modality relevance to select high-quality synthesized data, while Weighted Low-Rank Adaptation (WoRA) enables lightweight fine-tuning of a minimal set of model parameters.

- Extensive experiments on three widely-used benchmarks verify that our method could save 19.82% training time compared with the vanilla baseline, while also achieving competitive 76.37%, 66.65%, 67.90% Recall@1 accuracy on CUHK-PEDES, RSTPReid and ICFG-PEDES, respectively.

2 Related Work

Vision-Language Pre-training. Current VLP research predominantly bifurcates into coarse-grained and fine-grained methodologies. Coarse-grained approaches employ convolutional networks [23, 24, 26, 80] or visual Transformers [27, 28, 34, 45, 49] to extract and encode holistic image features, thereby constructing vision-language models (VLMs). Techniques such as SOHO [23] propose leveraging a Visual Dictionary (VD) for the extraction of comprehensive yet compact image features, facilitating enhanced cross-modal comprehension. ALBEF [34] introduces a contrastive loss for aligning image and text representations before their fusion through cross-modal attention, fostering more grounded vision-language representation learning. Additionally, it utilizes momentum distillation to augment learning capabilities from noisy network data. Although these holistic image-focused methods are efficient, their performance is generally outpaced by fine-grained approaches. Inspired by advancements in the NLP domain, fine-grained methods [10, 15, 31, 35, 40, 47, 60] employ pre-trained object detectors [3, 54] trained on annotated datasets of common objects, such as COCO [41] and Visual Genome [29]. This enables models to recognize and classify all potential object regions within images, representing them as a collection of object-centered features. For instance, VinVL [79] enhances visual representations for V+L tasks and develops an improved object detection model to provide object-centered image representations. However, this object-centered feature representation struggles to capture relationships between multiple objects across different regions, limiting its effectiveness in encoding multi-granularity visual concepts. Another limitation is the inability of the object detector to recognize uncommon objects not present in the training data. Recently, novel approaches have emerged to bridge the learning of object-level and image-level alignments. E2E-VLP [73] employs DETR [6] as the object detection module to enhance detection capabilities. KD-VLP [44] relies on external object detectors for object knowledge distillation, facilitating cross-modal alignment learning across different semantic layers. OFA [62] formulates visual-linguistic tasks as a sequence-to-sequence (seq2seq) problem, adhering to instruction-based learning during both pre-training and fine-tuning phases, eliminating the need for extra task-specific layers. Uni-Perceiver [90] constructs a unified perception architecture, using a single Transformer and shared parameters for diverse modes and tasks, employing a non-mixed sampling strategy for stable multi-task learning. X-VLM [77] and X2-VLM [78] propose an integrated model with a flexible modular architecture to simultaneously learn multi-granularity alignment and localization, achieving the capability to learn infinite visual concepts related to various text descriptions. In this work, we leverage the proficient vision-language pre-trained model to filter noisy data.

Text-Image Person Search. Based on the challenging task of language-based person search, which is a fine-grained, cross-modal retrieval challenge, a significant number of methodologies have been developed in recent years to tackle this issue. Existing approaches can broadly be classified into two categories: those based on cross-modal attention interaction [37, 57, 58, 67] and those without cross-modal attention interaction [9, 13, 66, 75, 76, 86]. Methods

leveraging cross-modal attention interactions facilitate correspondences between image regions and textual phrases by pairing inputs and predicting image-text matching scores through attention mechanisms. This enhances interaction between modalities, effectively bridging the modality gap, albeit at the cost of increased computational complexity. For instance, Li *et al.* [37] propose a recurrent neural network with gated neural attention to enhance cross-modal learning. Shao *et al.* [57] introduce a multimodal shared dictionary (MSD) to reconstruct visual and textual features, employing shared, learnable archetypes as queries. To improve person search performance, features for both modalities are extracted in the feature space with uniform granularity, ensuring semantic consistency. Conversely, methods without cross-modal attention interaction, through the construction of diverse model structures and objective functions, align representations of the two modalities within a shared feature space. These lightweight models, not reliant on complex cross-modal interactions, are computationally more efficient and have even achieved better results than their attention-based counterparts. Zheng *et al.* [86] build an end-to-end dual-path convolutional network to learn image and text representations to take full advantage of supervision capabilities. However, all the person search methods mentioned above fine-tune the entire network for high accuracy, which inherently slows down the process. In contrast, we propose a method based on cross-modal feature extraction for rapid candidate selection and data ranking scores. This approach aims to mitigate the impact of low-quality text-image pairs while reducing model computational parameters to maintain high efficiency in processing.

Data-Centric Learning. With the surge in popularity of large language models, an increasing demand for vast datasets for model training has emerged [52, 76–78]. However, open-source datasets constructed for training models in real-world scenarios, such as the MALS dataset [76], may encounter issues like incorrect text descriptions, poor image or text quality, and insufficient feature matching between image-text pairs, all of which can adversely affect model training performance. As the size of datasets expands, it is observed that their quality does not invariably scale in tandem [87]. Frequently, a subset of high-caliber data can attain or even exceed the utility of a voluminous but qualitatively inferior dataset. This phenomenon underscores the paramount importance of meticulously curating high-quality datasets, thereby highlighting the necessity for efficiency and precision in dataset construction. For instance, data selection methods [22, 51, 71] aim to identify and train only with the most relevant and informative examples, discarding irrelevant or redundant data. This leads to more efficient learning and improved model performance, especially when dealing with large datasets. Active learning seeks to reduce labeling costs by selecting the most informative instances for annotation. Data cleaning and preprocessing techniques aim to remove noise, errors, and inconsistencies from the data, making it more suitable for learning. Coreset selection [32, 33] focuses on identifying a small subset of data points (a core set) sufficient for training a model that performs well across the entire dataset. By selecting a representative subset, coreset selection can reduce the computational costs of training and improve model generalization. Our approach advocates for a coreset method aimed at enhancing model performance, proposing the deploy of the off-the-shelf visual-language models

to segment and filter the dataset for text and image pair matching scores, thereby obtaining a coreset dataset for effective training.

3 Methodology

3.1 Baseline Revisit

We do not pursue the network contribution in this work, but focus on the training efficiency. Our method can be adopted to most existing works. Without loss of generality, we apply the widely-used baseline, APTM [76], to simplify the illustration as well as a fair comparison with other methods. In particular, the framework comprises three encoders, *i.e.*, image encoder, text encoder, and cross encoder, along with two MLPs-based headers. The entire training process contains two phases, *i.e.*, pre-training on the synthesized dataset and fine-tuning on the downstream datasets. As shown in Figure 2, the [CLS] embedding represents the aggregated feature of the image / text from the image encoder and the text encoder respectively. The cross encoder integrates image and text representations for prediction tasks, leveraging the latter 6 layers of Bert to process the previously obtained text and image embeddings, thereby discerning their semantic relationship. We adopt two types of loss functions to bolster alignment constraints, tailored for both image-text and image-attribute associations. The image-text functions encompass Image-Text Contrastive Learning (ITC), Image-Text Matching Learning (ITM), and Masked Language Modeling (MLM), while Attribute Prompt Learning contains Image-Attribute Contrastive Learning (IAC) loss, Image-Attribute Matching Learning (IAM) Loss, and Masked Attribute Modeling (MAM) Loss. The overall APL loss is $\mathcal{L}_{APL} = \frac{1}{3} (\mathcal{L}_{iac} + \mathcal{L}_{iam} + \mathcal{L}_{mam})$, and the full pre-training loss is formulated as: $\mathcal{L}_{total} = \mathcal{L}_{itc} + \mathcal{L}_{itm} + \mathcal{L}_{mlm} + \eta \mathcal{L}_{APL}$, where η is empirically set as 0.8 following [76].

3.2 Data Filtering

As highlighted in our introduction, the acquisition of images is challenged by high annotation costs and concerns over individual privacy and security, necessitating the generation of a large volume of image-text pairs (I, T) through diffusion models, complemented by text descriptions generated by large language models [76]. Despite the potential for achieving high accuracy through extensive pretraining on such data, it becomes apparent that not all generated data are equally effective, with a significant portion being redundant. Additionally, these synthesized datasets often include noise in the form of image-text pairs with poor matching quality, which can inadequately represent the visual content of images. Such pairs serve as suboptimal signals for learning fine-grained visual-language alignment, potentially disrupting model training. This observation led us to ponder whether reducing the volume of data, focusing solely on a core set, could suffice for model training. Consequently, we devised a data filtering solution a data filtering approach based on the off-the-shelf large cross-modality model, *e.g.*, BLIP-2 [32], leveraging its efficient multimodal feature extraction capabilities to isolate features from both text and images (see Figure 3). We intend to find the appropriate filtering methodology and thresholds to isolate a high-quality core set.

Our method is executed in two phases, starting with the filtering of the synthesized dataset, *i.e.*, MALS, used for pretraining the model. Initially, all image-text pairs (I_M, T_M) within the dataset
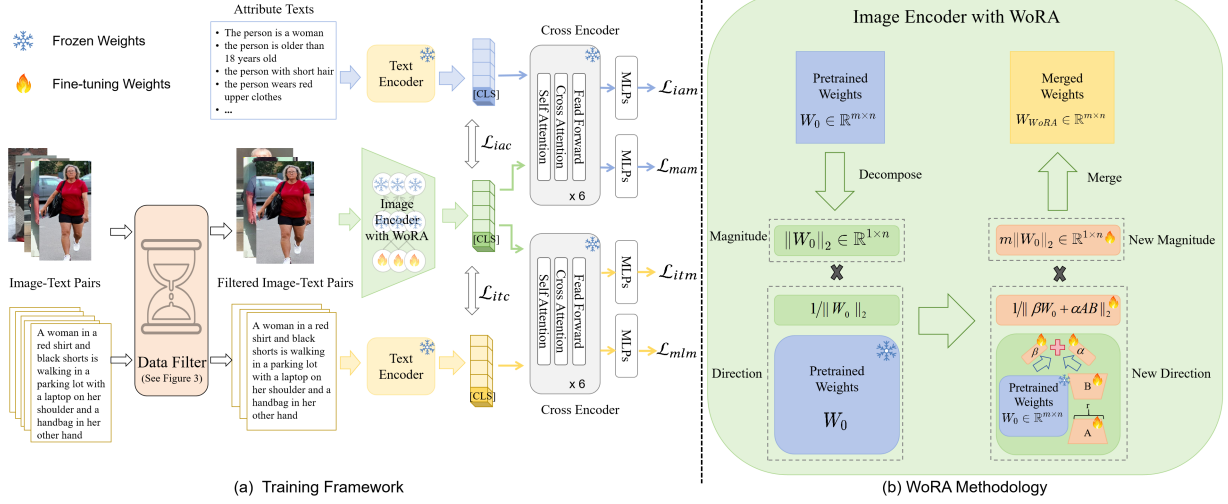
Figure 2: An overview of our framework. (a) shows the flow chart of the entire training pipeline. We first obtain the filtered training image-text pairs. Then we augment the text as attribute texts according to keywords. We extract the corresponding features through image encoder, text encoder and cross encoder. There are six loss objectives of both text-image and attribute-image matching tasks. (b) is an in-depth illustration of WoRA methodology, meticulously applied within the context of an image encoder. The model is updated by fine-tuning the decomposition of the pre-trained weights into amplitude and direction components and updating both components using LoRA [21] while adding the α and β learnable parameters. Since the image encoder consumes most GPU memory and time. In practice, we mainly apply the WoRA on the image encoder.
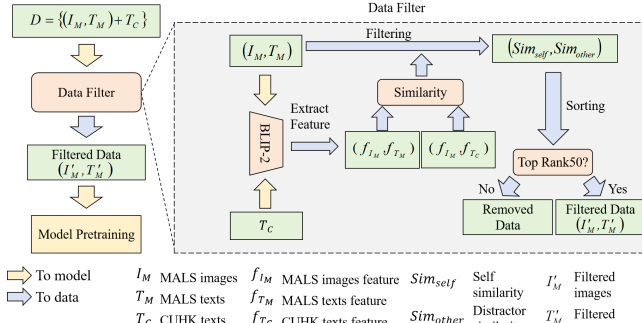

Figure 3: An overview of our data filtering process. We first employ Blip-2 [32] to extract features from the input image-text pair (I, T) and the distractor text T_C. Next, we compute the similarity and rank the results accordingly, ultimately generating the filtered dataset.

are processed. Given an image-text pair, the large cross-modality model, *i.e.*, BLIP-2, is employed to extract the corresponding image feature f_{I_M} and text feature f_{T_M}. Subsequently, the training set captions from the real-world downstream dataset, *i.e.*, CUHK-PEDES, acting as distractor texts T_C, are utilized. For each pair, we calculate the cosine similarity between projected image embeddings and projected text embeddings to ascertain their self-similarity Sim_{self}. Simultaneously, the distractor similarity Sim_{other} is determined by calculating the similarity between the image feature f_{I_M} and a randomly selected subset of 10,000 distractor texts T_C. The following is the formula for calculating the similarity:

$$Sim_{self} = cos\left(f_{I_M}, f_{T_M}\right), Sim_{other} = cos\left(f_{I_M}, f_{T_C}\right), \quad (1)$$

where $cos\left(,\right)$ denotes the cosine similarity. After computing these two types of similarities for all image-text pairs within the training

dataset, and then we sort texts according to the similarity from highest to lowest, we set a ranking threshold as 50. It means that an image-text pair is retained only if its self-similarity ranks within the top 50 of all calculated similarities. Through this threshold-based filtering, we discarded 21% of low-quality image-text pairs, retaining 79% to form a new pretraining dataset, Filtered-MALS.

Similarly, we also could filter the CUHK-PEDES training set in the fine-tuning phase to remove the noisy pair in the real-world training set. Employing the same calculation as in the first phase, we extract image features f_{I_C} and text features f_{T_C} for all pairs (I_C, T_C). Each image calculates the cosine similarity with both the ground-truth text and a subset of 10,000 random distractor texts $f_{T_{C'}}$. Similarly, the similarity can be formulated as:

$$Sim_{self} = cos\left(f_{I_C}, f_{T_C}\right), Sim_{other} = cos\left(f_{I_C}, f_{T_{C'}}\right). \quad (2)$$

Upon computing and ranking the similarities across the entire training set, we set a relatively loose ranking threshold of 1800, given that most pairs are human-annotated. This selection process results in the removal of 10% low-quality image-text pairs, leaving 90% to form a refined dataset, Refined-CUHK, for model fine-tuning.

Discussion. The mechanism of filtering. The motivation is within the spectrum of generated datasets. (1) It is worth noting that not all contents bear relevance, with a portion comprising low-quality image-text pairs. Such instances of subpar alignment between textual descriptions and corresponding images can invariably exert a deleterious impact, compromising the model training. Predominantly, the segments of the dataset that contribute most significantly to the performance of the model are those encompassing high-quality data, often referred to as the 'coreset'. (2) The extensive volume of data implicated in the pretraining phase incurs substantial computational costs, necessitating considerable
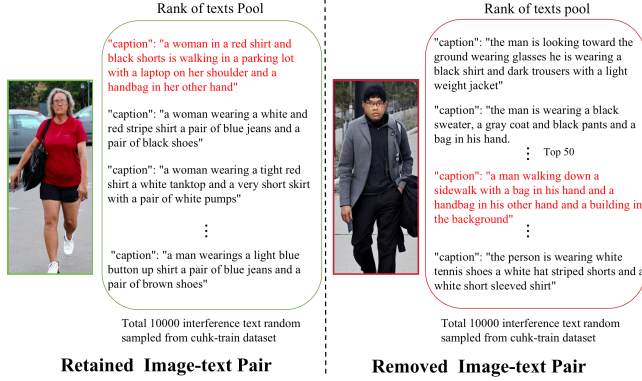
Figure 4: Visual explanation of data filtering. The part on the left of the image shows the high-quality image retained after our screening strategy and its corresponding red text description, while the person image on the right represents the low-quality image text pairs that are filtered out beyond the threshold, *i.e.*, top50. We deploy the real-world training set as distractors to filter low-relevance synthesized image-text pairs according to the similarity since there are no overlaps.
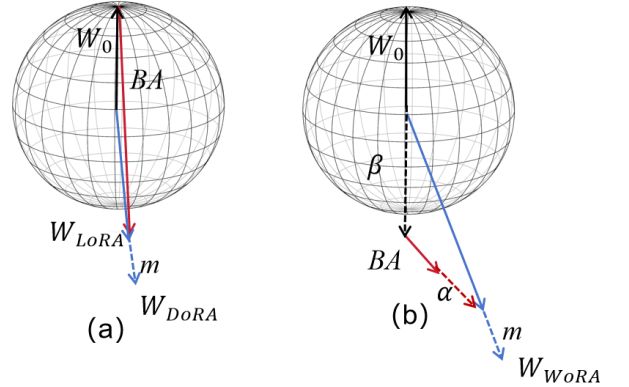

Figure 5: Intuitive comparison of LoRA, DoRA, and our proposed WoRA. (a) Here we show a common case during optimization, *i.e.*, negative correlation against W_0, which both LoRA and DoRA are struggling with. The bias parameter BA is hard to learn, considering the weight decay and other regularization. (b) In contrast, we deploy two float scalars in WoRA, *i.e.*, α and β, which could efficiently adjust the vector and provide better flexibility.

resources in terms of computational power and temporal investment. Thus, data filtering provides two primary advantages: (1) Our filtering algorithm identifies the core set of image-text pairs that are most crucial for improving data quality during model training. This selective process ensures the model is exposed to higher-quality data, enhancing its learning capabilities. (2) Retaining only the core set significantly reduces the dataset size, thereby decreasing computational overhead during training. This results in reduced computational requirements and training time. Figure 4 illustrates an example of the data filtering process.

3.3 Weighted Low-Rank Adaptation

While the introduction of a pretrain-finetune paradigm to train models for person search tasks has achieved commendable results [76], the expansion in model and dataset sizes significantly increases the number of parameters to be trained, demanding substantial computational resources and sacrificing training efficiency. To address this issue, several parameter-efficient fine-tuning methods [20] have been proposed, aiming to fine-tune pretrained models using the minimum number of parameters. Among these, LoRA [21] has gained popularity due to its simplicity and efficacy. LoRA employs a low-rank decomposition for the pretrained weight matrix $W_0 \in \mathbb{R}^{m \times n}$, $W_0 + \triangle W = W_0 + BA$, where $B \in \mathbb{R}^{m \times r}$, $A \in \mathbb{R}^{r \times n}$, rank $r \ll \min(m, n)$. $\triangle W$ is adjusted by $\frac{8}{r}$. Inspired by the LoRA [21], which updates only a small part of the model weight to improve efficiency, DoRA [43] decomposes the weight into two parts: direction and amplitude. DoRA improves the adaptability and efficiency of the model, which can be formulated as $W_{DoRA} = m \frac{W_0 + BA}{\|W_0 + BA\|_2}$, where $m \in \mathbb{R}^{1 \times n}$ is the magnitude vector. $\|\cdot\|_2$ denotes the L2 norm of a matrix across each column.

However, these two methods still limit the parameter freedom. Therefore, we introduce the Weighted Low-Rank Adaptation (WoRA) model to address the capacity gap still present between LoRA [21] and fine-tuning (see Figure 2 right). Drawing from the DoRA [43] approach, which reparameterizes model weights into magnitude

and direction components for fine-tuning, WoRA introduces new learnable parameters, α and β, to facilitate parameter learning. Given that pretrained weights already possess a vast repository of knowledge suitable for various downstream tasks, we configure these learnable parameters to ensure the model acquires sufficient capability in both magnitude and direction. This allows for the model to adapt to downstream tasks by updating parameters that exhibit significant magnitude or directional changes. The formal representation of our WoRA model is:

$$W_{WoRA} = m \frac{\beta * W_0 + \alpha * BA}{\|\beta * W_0 + \alpha * BA\|_2}. \tag{3}$$

In our model, parameters denoted with an overline represent trainable parameters. The proposed Weighted Low-Rank Adaptation (WoRA) demonstrates learning capabilities comparable to full fine-tuning. During inference, WoRA integrates with pretrained weights, introducing no additional latency while enhancing both learnability and computational efficiency. Moreover, by incorporating new learnable parameters, WoRA achieves superior performance compared to LoRA [21] and DoRA [43].

Discussion. Why is WoRA better than LoRA and DoRA? Both DoRA and LoRA are specific cases of the proposed WoRA. Specifically, DoRA can be derived from WoRA by setting $\alpha = 8$ and $\beta = 1$ as fixed constants. Similarly, setting $\alpha = 8$, $\beta = 1$, and $m = 1$ yields the identical behavior as LoRA. Consider a scenario involving negative correlation, which is common during training. As illustrated in Figure 5 (a), both LoRA and DoRA struggle with the typical negative correlation against the original weight W_0, necessitating **the learning of a large BA—a challenging and unstable process due to weight decay and other regularization terms**. In contrast, by introducing two scalar parameters, α and β, WoRA provides a more flexible and adaptive space for optimization. Our method can readily achieve negative correlation by setting a negative β, as depicted in Figure 5 (b). Additionally, α offers fine-grained control over BA during optimization. We will add an illustration of this to the revised version. **Time Cost**: WoRA introduces three

additional trainable parameters compared to LoRA, resulting in a slightly longer training time due to the increased complexity of weight calculation. Although WoRA's training time is marginally higher than that of DoRA and LoRA, the additional trainable parameters enable better weight adjustment, ultimately enhancing performance. As shown in Table 5, a small increase in training time (approximately 30 minutes) significantly improves the model's fine-tuning capability. **Memory Cost**: It is noteworthy that our space complexity is equivalent to DoRA, while WoRA introduces greater freedom in direction and magnitude, thereby facilitating fine-tuning. WoRA adds only two constant learnable scalars per weight update compared to DoRA, resulting in a negligible increase in space complexity.

4 Experiment

4.1 Experimental Setup

Datasets. Our study involves processing and training on four datasets. For pretraining, we utilize the synthetic dataset, MALS [76], which comprises 1,510,330 image-text pairs, each annotated with relevant attribute labels. For fine-tuning, we employ CUHK-PEDES [37], RSTPReid [89], and ICFG-PEDES. CUHK-PEDES aggregates 40,206 images of 13,003 individuals from 5 existing person reID datasets: CUHK03 [39], Market-1501 [83], SSM [72], VIPER [17], and CUHK01 [38]. Each image is paired with descriptions from two sentences, resulting in a total of 80,412 sentences. Our approach is evaluated on the public text-based person search dataset CUHK-PEDES. RSTPReid contains 20,505 images of 4,101 individuals, compiled from MSMT17 [70]. ICFG-PEDES [13], also derived from MSMT17, consists of 54,522 images of 4,102 individuals.

Evaluation metrics. We adopt the mean Average Precision (AP) and Recall@1,5,10 as our primary evaluation metrics. The Recall@K, whose value is 1 if the first matched image has appeared before the K-th image. Recall@K is sensitive to the position of the first matched image and suits the test set with only one true-matched image in the gallery. The average precision (AP) is the area under the PR(Precision-Recall) curve, considering all ground-truth images in the gallery. mAP is calculated and averaged for the average accuracy (AP) of each category.

Implementation Details. Our model with the proposed WoRA, undergoes pretraining across 32 epochs on 8 NVIDIA A800 GPUs via Pytorch, adopting a batch size of 150. The optimization strategy employs the AdamW optimizer [46], integrating a weight decay factor of 0.01. Initiation of the learning rate is set at $1e^{-5}$, incorporating a warm-up phase over the initial 2600 steps, which then transitions into a linear decay schedule ranging from $1e^{-4}$ down to $1e^{-5}$. Image preprocessing includes resizing to 384 × 128 dimensions, coupled with augmentation strategies such as random horizontal flipping, RandAugment [11], and random erasing [88]. For the pretraining phase, the image encoder is initialized using Swin-B [45], enhanced by the application of the WoRA method. Both the text encoder and cross encoder commence with configurations derived from the initial and final 6 layers of Bert [12], respectively. Subsequent to the pretraining, the model is fine-tuned on designated downstream datasets over 30 epochs. Initial WoRA settings for the image encoder, rank = 8, α = 8, and β = 1, are preserved, with the learning rate

Method	#Parameter	R@1	R@5	R@10	mAP
CNN-RNN [53]	-	8.07	-	32.47	-
GNA-RNN [37]	-	19.05	-	53.64	-
PWM-ATH [8]	-	27.14	49.45	61.02	-
GLA [7]	-	43.58	66.93	76.2	-
Dual Path [86]	-	44.40	66.26	75.07	-
CMPM+CMPC [81]	-	49.37	-	79.21	-
MIA [48]	-	53.10	75.00	82.90	-
A-GANet [42]	-	53.14	74.03	81.95	-
ViTAA [65]	177M	55.97	75.84	83.52	51.60
IMG-Net [68]	-	56.48	76.89	85.01	-
CMAAM [1]	-	56.68	77.18	84.86	-
HGAN [82]	-	59.00	79.49	86.62	-
NAFS [16]	189M	59.94	79.86	86.70	54.07
DSSL [89]	-	59.98	80.41	87.56	-
MGEL [61]	-	60.27	80.01	86.74	-
SSAN [13]	-	61.37	80.15	86.73	-
NAFS [16]	189M	61.50	81.19	87.51	-
TBPS [18]	43M	61.65	80.98	86.78	-
TIPCB [9]	185M	63.63	82.82	89.01	-
LBUL [67]	-	64.04	82.66	87.22	-
CAIBC [66]	-	64.43	82.87	88.37	-
AXM-Net [14]	-	64.44	80.52	86.77	58.73
LGUR [57]	-	65.25	83.12	89.00	-
CFine [74]	-	69.57	85.93	91.15	-
VGSG [19]	-	71.38	86.75	91.86	67.91
TBPS-CLIP [5]	149M	73.54	88.19	92.35	65.38
IRRA [25]	194M	73.38	89.93	93.71	66.13
RaSa [4]	210M	76.51	90.29	94.25	69.38
APTM [76]	214M	76.53	90.04	94.15	66.91
Baseline*	214M	75.42	88.86	92.77	66.61
Ours	**127M**	**76.38**	**89.72**	**93.49**	**67.22**

Table 1: Performance Comparison on CUHK-PEDES. Here we show the performance of the previous methods on the recall@1,5,10, mAP in % and the parameter number. Baseline*: We re-implement APTM [76].

commencing at $1e^{-4}$. This phase includes a warm-up period spanning the first three epochs, succeeded by a methodical reduction of the learning rate according to a linear scheduler. In addition to image data augmentation mentioned during pretraining, Easy Data Augmentation (EDA) [69] is employed for text data augmentation, with the batch size as 120. For each text query, its cosine similarity with all images is calculated, selecting the top 128 candidate images. Subsequently, the match probability between the text query and each selected image candidate is computed and ranked [76].

4.2 Comparison with Competitive Methods

We deploy the Weighted Low-Rank Adaptation (WoRA) for text-based person search tasks. Performance comparisons are made using Recall@1,5,10 and mean Average Precision (mAP) metrics, alongside a comparison of parameter count (params in Millions, M) and computational efficiency (FLOPs) against the baseline model APTM. Through trainable adjustments to weight parameters, the WoRA method indicated robust performance across both datasets, significantly reducing computational parameters and time. Specifically, compare to APTM, which is trained on 1.51M data, our data filtering algorithm remove 21% of low-quality, noisy data, utilizing 1.19M data for computations. Our implementation of WoRA on the CUHK-PEDES dataset reduce trainable parameters to 127.37M, a

Method	#Parameter	R@1	R@5	R@10	mAP
DSSL [89]	-	32.43	55.08	63.19	-
LBUL [67]	-	45.55	68.20	77.85	-
IVT [58]	-	46.70	70.00	78.80	-
CAIBC [66]	-	47.35	69.55	79.00	-
CFine [74]	-	50.55	72.50	81.60	-
TBPS-CLIP [5]	149M	61.95	83.55	88.75	48.26
IRRA [25]	194M	60.20	81.30	88.20	47.17
RaSa [4]	210M	66.90	86.50	91.35	52.31
APTM [76]	214M	67.50	85.70	91.45	52.56
Baseline*	214M	66.40	85.55	91.10	52.21
Ours	**127M**	**66.85**	**86.45**	**91.10**	**52.49**

Table 2: Performance Comparison on RSTPReid. Here we show the performance of the previous methods on the recall@1,5,10 and mAP in % and the parameter number. Baseline*: We re-implement APTM [76].

Method	#Parameter	R@1	R@5	R@10	mAP
Dual Path [86]	-	38.99	59.44	68.41	-
CMPM+CMPC [81]	-	43.51	65.44	74.26	-
MIA [48]	-	46.49	67.14	75.18	-
SCAN [30]	-	50.05	69.65	77.21	-
ViTAA [65]	177M	50.98	68.79	75.78	-
SSAN [13]	-	54.23	72.63	79.53	-
IVT [58]	-	56.04	73.60	80.22	-
LGUR [57]	-	59.02	75.32	81.56	-
CFine [74]	-	60.83	76.55	82.42	-
TBPS-CLIP [5]	149M	65.05	80.34	85.47	39.83
IRRA [25]	194M	63.46	80.25	85.82	38.06
RaSa [4]	210M	65.28	80.04	85.12	41.29
APTM [76]	214M	68.51	82.99	87.56	41.22
Baseline*	214M	67.81	82.70	87.32	41.22
Ours	**127M**	**68.35**	**83.10**	**87.53**	**42.60**

Table 3: Performance Comparison on ICFG-PEDES. Here we show the performance of the previous methods on the recall@1,5,10 and mAP in % and the parameter number. Baseline*: We re-implement APTM [76].

41% decrease from APTM, with FLOPs reduce to 23.21G, a 39% reduction. The overall training duration for pretraining and fine-tuning is cut by 19.82%, with our model achieving slight improvements in recall rates and mAP, as evidenced in Table 1. Moreover, the pretrained model adjusted through WoRA achieves competitive performance on the RSTPReid and ICFG-PEDES dataset, as shown in Table 2 and 3. To show the performance of our model more intuitively, we present three visual retrieval results on CUHK-PEDES in Figure 6. Our model adeptly captures fine-grained, word-level details, enabling it to accurately differentiate subtle variations in clothing colors among individuals. Furthermore, it exhibits robust retrieval capabilities, effectively identifying subjects even when parts of their details are obscured. This indicates the strong performance of our model in handling nuanced visual variations and partial occlusions within complex scenes.

4.3 Ablation Study and Further Discussion

The impact of data filtering. Here, "baseline" refers to the APTM method as implemented in our experimental setup, and "top50" denotes our data filtering approach with a threshold set to top50 for selecting the pre-training dataset MALS. From Table 4, it is evident that training the pre-trained models with our filtered dataset
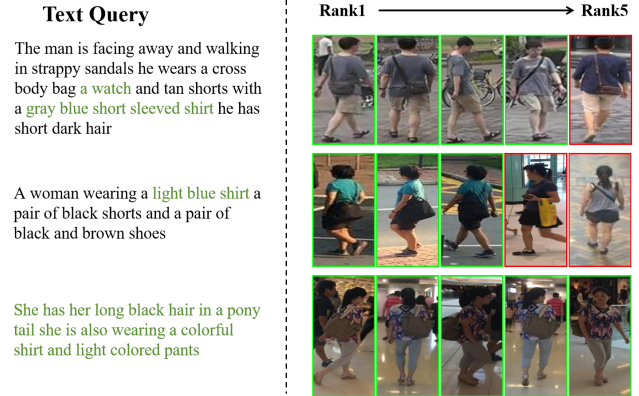

Figure 6: Qualitative person search results using text query of our methods, placing in descending order from left to right based on matching probability. The images in green boxes are the correct matches, and the images in red boxes are the wrong matches. The green texts show that our results successfully match.
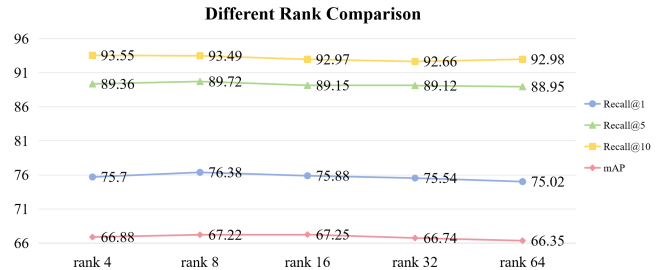

Figure 7: The impact of different WoRA ranks on performance. We observe that the result is not sensitive to the rank. Generally, rank=8 is the best hyper-parameter in terms of performance on Recall@1,5,10.

results in improvements of 6.26% in Recall@1, demonstrating the efficacy of our dataset filtering for pre-training. Similarly, we find that filtering on the downstream dataset also facilitates the learning, since the filtering algorithm generally removes the noise. In Table 5, the notation "ft90%" signifies the leverage of our dataset filtering method to refine the finetune dataset CUHK-PEDES, ultimately retaining 90% of its data for training. The model trained after applying "ft90%" exhibits a 0.25% increase in Recall@1, substantiating the effectiveness of dataset filtering on the comprehensive model.

Moreover, we assess the impact of Weighted Low-Rank Adaptation (WoRA) on the complete model. Initially, comparing the impacts of using LoRA [21] and DoRA [43] models of Low-Rank Adaptation on our fully trained model revealed improvements in model speed but not in performance. Subsequently, applying WoRA to models trained on the top50 pre-training selection and the ft90% finetuned dataset, we achieve a Recall@1 of 76.38%, surpassing the baseline by 0.96%, and an mAP of 67.22%, exceeding the baseline by 0.61%, thereby establishing state-of-the-art (SOTA) mAP performance. The computation time for the complete model is reduced from 23h to 18h, marking a 19.82% acceleration. In Figure 7, we compare the effects of different rank values on the performance of WoRA. The experiment shows that the model has the best performance on Recall@1,5,10 when rank is equal to 8. We can observe that model performance may not be as sensitive in different ranks.

Jintao Sun, Hao Fei, Gangyi Ding, and Zhedong Zheng

Method	# Data (M)	# Trainable Params (M)	Flops (G)	CUHK-PEDES R@1	mAP	RSTPReid R@1	mAP	ICFG-PEDES R@1	mAP	Time (hours)
Pretraining Stage										
Baseline* (APTM [76])	1.51M	213.99 M	38.02 G	3.99	3.62	4.40	3.95	0.77	0.59	18h
Ours (top50)	1.19M	213.99 M	38.00 G	10.25	10.24	12.65	9.37	8.07	2.35	14h
Ours (top50+WoRA)	**1.19M**	**127.37 M**	**23.21 G**	**10.71**	**10.33**	**13.00**	**9.59**	**10.80**	**3.10**	14h*
Finetune Stage										
Baseline* (APTM [76])	0.068M	213.99 M	44.93 G	75.42	66.61	66.32	52.30	67.66	41.98	4.2h
Ours (top50)	0.061M	213.99 M	44.93 G	75.67	66.27	66.40	52.21	67.80	42.38	3.8h
Ours (top50+WoRA)	**0.061M**	**127.37 M**	**30.13 G**	**76.38**	**67.22**	**66.85**	**52.49**	**68.35**	**42.60**	3.8h*

Table 4: Compared with APTM method at recall@1 and mAP results on CUHK-PEDES, RSTPReid and ICFG-PEDES. Meanwhile, we also compare the data volume, params (M) and Flops (G) of the model. Ablation study about our methods on pretrain. The top50 denotes the results trained by using a pre-trained dataset filtered using a data filtering method. The top50+WoRA is our ultimate two-stage approach, and by adding WoRA to fine-tune the model, we can improve the performance of the pre-trained model while saving 19.82% training time. Time* indicates that the time decrease significantly because the bottleneck of model at other places. Thus, even though we optimize the GPU FPS, it have minimal impact on the overall computation time.

Method	top50	ft 90%	R@1	mAP
Baseline*			75.42	66.61
Baseline*	✓		75.83	66.70
Baseline*	✓	✓	75.67	66.27
LoRA	✓		74.40	64.95
LoRA	✓	✓	74.29 (-1.38)	65.59 (-0.68)
DoRA	✓		75.49	66.92
DoRA	✓	✓	75.73 (+0.06)	66.75 (+0.48)
WoRA (Ours)	✓		75.67	67.09
WoRA (Ours)	✓	✓	76.38 (+0.71)	67.22 (+0.95)

Table 5: Comparison of our WoRA with baseline, LoRA and DoRA in the different situations. Finally, the experiment shows that our methods achieve the best recall@1 and mAP in %. ft 90% denotes the utilization of our dataset filtering method to refine the finetune dataset CUHK-PEDES, ultimately retaining 90% of its data for training, while top50 is using a data filtering approach with a threshold set to top50 for selecting the pre-training dataset MALS.

Performance can be improved basically with the WoRA model. The results we present in this paper are obtained using rank=8, which has achieved the best performance on Recall@1,5,10. As shown in Table 5, we could observe two points (1) WoRA is better than both LoRA and DoRA, whether the performance of Recall@1 or mAP. (2) Furthermore, if we both leverage the top50 and ft90% filtering strategy, WoRA is also the best performing, which surpasses the baseline by 0.71% Recall@1 and 0.95% mAP under the same conditions. Meanwhile, in order to compare the effectiveness of WoRA. We control all the learning rates unchanged and compare the performance of LoRA, DoRA, and WoRA (ous) in Figure 8. we use the same ranks for LoRA, DoRA, and WoRA for comparison. We can clearly see from the figure that under the condition of a fixed learning rate, the WoRA method proposed by us is consistently better than the LoRA and DoRA model mAP under the condition of the same rank. For example, when the same rank is 64, our method is +0.57% mAP higher than LoRA. +2.38% mAP higher than DoRA. Therefore, it can be clearly seen that our WoRA method is superior to DoRA and LoRA under the condition of a fixed learning rate. **The impact of WoRA.** We further investigate the impact of Low-Rank Adaptation on pre-training, where "WoRA" in Table 4 and 5
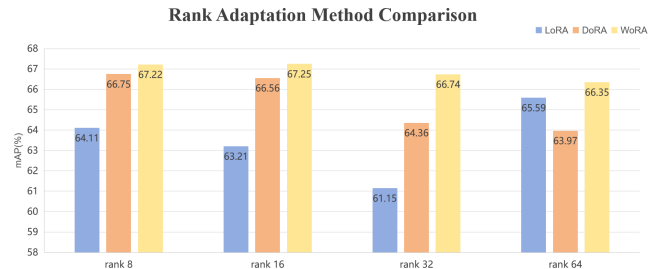

Figure 8: Compare the performance of LoRA, DoRA, and WoRA in terms of different rank. Our WoRA consistently surpasses both LoRA and DoRA on various rank settings.

signifies our Weighted Low-Rank Adaptation approach. Implementing WoRA in the pre-trained models led to a +6.72% boost in Recall@1 and a +6.71% enhancement in mAP, concurrently reducing the time by 22.22%, thereby validating the efficiency of our dataset filtering and WoRA in model pre-training.

5 Conclusion

In this work, we introduce a new Filtering-WoRA paradigm, which contains a filtering algorithm to identify this crucial data subset and WoRA layers (Weighted Low-Rank Adaptation) for light fine-tuning. Filtering strategy for image-text pairs within language-based person search datasets, designed to isolate a core set from large-scale, noise-containing datasets of generated image-text pairs. WoRA (Weighted Low-Rank Adaptation) learning strategy to efficiently update the portion of model parameters. Extensive experiments indicate that our approach is 19.82% faster than existing language-based person search methods while maintaining comparable accuracy with state-of-the-art (SOTA) language-based person search models. On three public benchmarks, CUHK-PEDES, RSTPReeid, and ICFG-PEDES, our method achieves competitive recall rates and mean Average Precision (mAP).

Acknowledgements

References

[1] Surbhi Aggarwal, R. Venkatesh Babu, and Anirban Chakraborty. 2020. Text-based Person Search via Attribute-aided Matching. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2606–2614. doi:10.1109/WACV45572.2020.9093640

[2] Shafiq Ahmad, Pietro Morerio, and Alessio Del Bue. 2023. Person Re-Identification without Identification via Event Anonymization. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 11098–11107. doi:10.1109/ICCV51070.2023.01022

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2018.00636

[4] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. 2023. RaSa: Relation and Sensitivity Aware Representation Learning for Text-based Person Search. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-2023)*. International Joint Conferences on Artificial Intelligence Organization. doi:10.24963/ijcai.2023/62

[5] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. 2023. An Empirical Study of CLIP for Text-based Person Search. *arXiv preprint arXiv:2308.10045* (2023).

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. *End-to-End Object Detection with Transformers*. 213–229. doi:10.1007/978-3-030-58452-8_13

[7] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. 2018. Improving Deep Visual Representation for Person Re-identification by Global and Local Image-language Association. In *Computer Vision − ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI* (Munich, Germany). Springer-Verlag, Berlin, Heidelberg, 56–73. doi:10.1007/978-3-030-01270-0_4

[8] Tianlang Chen, Chenliang Xu, and Jiebo Luo. 2018. Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1879–1887. doi:10.1109/WACV.2018.00208

[9] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. 2022. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing* (Jul 2022), 171–181. doi:10.1016/j.neucom.2022.04.081

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. *UNITER: UNiversal Image-TExt Representation Learning*. 104–120. doi:10.1007/978-3-030-58577-8_7

[11] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. doi:10.1109/cvprw50498.2020.00359

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[13] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. 2021. Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition* (Jul 2021).

[14] Ammarah Farooq, Muhammad Awais, Josef Kittler, and Syed Safwan Khalid. 2022. AXM-Net: Implicit Cross-Modal Feature Alignment for Person Re-identification. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 4 (Jun. 2022), 4477–4485. doi:10.1609/aaai.v36i4.20370

[15] Zhe Gan, Yen-Chun Chen, Pingqing Fu, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-Scale Adversarial Training for Vision-and-Language Representation Learning. *Neural Information Processing Systems,Neural Information Processing Systems* (Jun 2020).

[16] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. 2021. Contextual Non-Local Alignment over Full-Scale Representation for Text-Based Person Search. arXiv:2101.03036 [cs.CV]

[17] DouglasA. Gray, Shane Brennan, and Hai Tao. 2007. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. (Jan 2007).

[18] Xiaoping Han, Sen He, Li Zhang, and Tao Xiang. 2021. Text-Based Person Search with Limited Data. In *British Machine Vision Conference*. https://api.semanticscholar.org/CorpusID:239050116

[19] Shuting He, Hao Luo, Wei Jiang, Xudong Jiang, and Henghui Ding. 2024. VGSG: Vision-Guided Semantic-Group Network for Text-Based Person Search. *IEEE Transactions on Image Processing* 33 (2024), 163–176. doi:10.1109/TIP.2023.3337653

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. arXiv:1902.00751 [cs.LG]

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[22] Xu Huang, Jianxun Lian, Hao Wang, Hao Liao, Defu Lian, and Xing Xie. 2024. A Data-Centric Multi-Objective Learning Framework for Responsible Recommendation Systems. In *Proceedings of the ACM on Web Conference 2024*. 3129–3138.

[23] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning. 12971–12980. doi:10.1109/CVPR46437.2021.01278

[24] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition* (Apr 2020).

[25] Ding Jiang and Mang Ye. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. 2020. In Defense of Grid Features for Visual Question Answering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr42600.2020.01028

[27] Zhigang Kan, Liwen Peng, Yifu Gao, Ning Liu, Linbo Qiao, and Dongsheng Li. 2024. LFDe: A Lighter, Faster and More Data-Efficient Pre-training Framework for Event Extraction. In *Proceedings of the ACM on Web Conference 2024*. 3964–3975.

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*. PMLR, 5583–5594.

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* (May 2017), 32–73. doi:10.1007/s11263-016-0981-7

[30] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*. 201–216.

[31] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence* (Jun 2020), 11336–11344. doi:10.1609/aaai.v34i07.6795

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (, Honolulu, Hawaii, USA,) *(ICML'23)*. JMLR.org, Article 814, 13 pages.

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 12888–12900. https://proceedings.mlr.press/v162/li22n.html

[34] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[35] LiunianHarold Li, Mark Yatskar, Dong Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *Cornell University - arXiv,Cornell University - arXiv* (Aug 2019).

[36] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE international conference on computer vision*. 1890–1899.

[37] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person Search with Natural Language Description. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2017.551

[38] Wei Li, Rui Zhao, and Xiaogang Wang. 2013. Human reidentification with transferred metric learning. In *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11*. Springer, 31–44.

[39] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/cvpr.2014.27

[40] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks*. 121–137. doi:10.1007/978-3-030-58577-8_8

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. *Microsoft COCO: Common Objects in Context*. 740–755. doi:10.1007/978-3-319-10602-1_48

[42] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. 2019. Deep Adversarial Graph Attention Convolution Network for Text-Based

Person Search. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 665–673. doi:10.1145/3343031.3350991

[43] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 1299, 22 pages.

[44] Yongfei Liu, Chenfei Wu, Shao-Yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. 2022. KD-VLP: Improving End-to-End Vision-and-Language Pretraining with Object Knowledge Distillation. In *Findings of the Association for Computational Linguistics: NAACL 2022*. doi:10.18653/v1/2022.findings-naacl.119

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. doi:10.1109/iccv48922.2021.00986

[46] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:53592270

[47] Jing Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Neural Information Processing Systems,Neural Information Processing Systems* (Aug 2019).

[48] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. 2019. Improving Description-Based Person Re-Identification by Multi-Granularity Image-Text Alignments. *IEEE Transactions on Image Processing* 29 (2019), 5542–5556. https://api.semanticscholar.org/CorpusID:195345251

[49] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers. *ArXiv* abs/2208.06366 (2022). https://api.semanticscholar.org/CorpusID:251554649

[50] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV]

[51] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash J. Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Kumar Mahajan. 2023. Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 6967–6977. https://api.semanticscholar.org/CorpusID:255522657

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[53] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 49–58.

[54] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Jun 2017), 1137–1149. doi:10.1109/tpami.2016.2577031

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10674–10685. doi:10.1109/CVPR52688.2022.01042

[56] Sasa Sambolek and Marina Ivasic-Kos. 2021. Automatic Person Detection in Search and Rescue Operations Using Deep CNN Detectors. *IEEE Access* 9 (2021), 37905–37922. doi:10.1109/ACCESS.2021.3063681

[57] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhi hao Lin, Jian Wang, and Changxing Ding. 2022. Learning Granularity-Unified Representations for Text-to-Image Person Re-identification. *Proceedings of the 30th ACM International Conference on Multimedia* (2022). https://api.semanticscholar.org/CorpusID:250627620

[58] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. 2022. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*. Springer, 624–641.

[59] A. Syed Musthafa, Dinesh S, Dinesh Kumar K, Jeeva C, and Madesh S. 2024. Digital Vigilance: AI Solutions in the Quest for Missing Persons using face recognition with deep learning algorithms. In *2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*. 1–6. doi:10.1109/AIMLA59606.2024.10531327

[60] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. doi:10.18653/v1/d19-1514

[61] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li. 2021. Text-based Person Search via Multi-Granularity Embedding Learning. In *International Joint Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:237100188

[62] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *CoRR* abs/2202.03052 (2022).

[63] Xiaodong Wang, Zhedong Zheng, Yang He, Fei Yan, Zhiqiang Zeng, and Yi Yang. 2022. Soft Person Reidentification Network Pruning via Blockwise Adjacent Filter Decaying. *IEEE Transactions on Cybernetics* 52, 12 (2022), 13293–13307. doi:10.1109/TCYB.2021.3130047

[64] Xiaodong Wang, Zhedong Zheng, Yang He, Fei Yan, Zhiqiang Zeng, and Yi Yang. 2023. Progressive Local Filter Pruning for Image Retrieval Acceleration. *IEEE Transactions on Multimedia* 25 (2023), 9597–9607. doi:10.1109/TMM.2023.3256092

[65] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. 2020. ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 402–420.

[66] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. 2022. CAIBC: Capturing All-round Information Beyond Color for Text-based Person Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*. doi:10.1145/3503161.3548057

[67] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. 2022. Look Before You Leap: Improving Text-based Person Retrieval by Learning A Consistent Cross-modal Common Manifold. In *Proceedings of the 30th ACM International Conference on Multimedia*. doi:10.1145/3503161.3548166

[68] Zijie Wang, Aichun Zhu, Zhe Zheng, Jing Jin, Zhouxin Xue, and Gang Hua. 2020. IMG-Net: inner-cross-modal attentional multigranular network for description-based person re-identification. *Journal of Electronic Imaging* 29 (2020), 043028 – 043028. https://api.semanticscholar.org/CorpusID:221380389

[69] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. doi:10.18653/v1/d19-1670

[70] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/cvpr.2018.00016

[71] Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. 2024. Vision-Language Dataset Distillation. arXiv:2308.07545 [cs.CV]

[72] Tong Xiao, Shuang Li, Bochao Wang, Lin Li, and Xiaogang Wang. 2016. End-to-End Deep Learning for Person Search. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition* (Apr 2016).

[73] Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. doi:10.18653/v1/2021.acl-long.42

[74] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. 2023. CLIP-Driven Fine-Grained Text-Image Person Re-Identification. *IEEE Transactions on Image Processing* 32 (2023), 6032–6046. doi:10.1109/TIP.2023.3327924

[75] Shuyu Yang, Yaxiong Wang, Li Zhu, and Zhedong Zheng. 2024. Beyond Walking: A Large-Scale Image-Text Benchmark for Text-based Person Anomaly Search. arXiv:2411.17776 [cs.CV] https://arxiv.org/abs/2411.17776

[76] Shuyu Yang, Yinan Zhou, Yaxiong Wang, Yujiao Wu, Li Zhu, and Zhedong Zheng. 2023. Towards Unified Text-based Person Retrieval: A Large-scale Multi-Attribute and Language Search Benchmark. In *Proceedings of the 2023 ACM on Multimedia Conference*.

[77] Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:244129883

[78] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022. X2-VLM: All-In-One Pre-trained Model For Vision-Language Tasks. *arXiv:2211.12402* (2022).

[79] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr46437.2021.00553

[80] Wen Zhang, Yushan Zhu, Mingyang Chen, Yuxia Geng, Yufeng Huang, Yajing Xu, Wenting Song, and Huajun Chen. 2023. Structure pretraining and prompt tuning for knowledge graph transfer. In *Proceedings of the ACM Web Conference 2023*. 2581–2590.

[81] Ying Zhang and Huchuan Lu. 2018. Deep Cross-Modal Projection Learning for Image-Text Matching. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 707–723.

[82] Kecheng Zheng, Wu Liu, Jiawei Liu, Zheng-Jun Zha, and Tao Mei. 2020. Hierarchical Gumbel Attention Network for Text-based Person Search. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 3441–3449.

doi:10.1145/3394171.3413864

[83] Liang Zheng, Liyue Shen, Lei Tian, Shengjin Wang, Jiahao Bu, and Qi Tian. 2015. Person Re-identification Meets Image Search. *Tpami* (Feb 2015).

[84] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. 2019. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2138–2147.

[85] Zhedong Zheng and Liang Zheng. 2024. Object Re-identification: Problems, Algorithms and Responsible Research Practice. In *The Boundaries of Data*. Amsterdam University Press.

[86] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-Path Convolutional Image-Text Embedding with Instance Loss. *ACM Transactions on Multimedia Computing, Communications, and Applications* (May 2020), 1–23. doi:10.1145/3383184

[87] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*. 3754–3762.

[88] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random Erasing Data Augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* (Jun 2020), 13001–13008. doi:10.1609/aaai.v34i07.7000

[89] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 209–217. doi:10.1145/3474085.3475369

[90] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 2022. Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16783–16794. doi:10.1109/CVPR52688.2022.01630