Mars: Paying More Attention to Visual Attributes for Text-based Person Search

ALEX ERGASTI* and TOMASO FONTANINI*, Department of Engineering and Architecture, University of Parma, IT
CLAUDIO FERRARI, Department of Engineering and Architecture, University of Parma, IT
MASSIMO BERTOZZI, Department of Engineering and Architecture, University of Parma, IT
ANDREA PRATI, Department of Engineering and Architecture, University of Parma, IT

Text-based person search (TBPS) is a problem that gained significant interest within the research community. The task is that of retrieving one or more images of a specific individual based on a textual description. The multi-modal nature of the task requires learning representations that bridge text and image data within a shared latent space. Existing TBPS systems face two major challenges. One is defined as *inter-identity* noise that is due to the inherent vagueness and imprecision of text descriptions and it indicates how descriptions of visual attributes can be generally associated to different people; the other is the *intra-identity* variations, which are all those nuisances *e.g.*, pose, illumination, that can alter the visual appearance of the same textual attributes for a given subject. To address these issues, this paper presents a novel TBPS architecture named MARS (Mae-Attribute-Relation-Sensitive), which enhances current state-of-the-art models by introducing two key components: a Visual Reconstruction Loss and an Attribute Loss. The former employs a Masked AutoEncoder trained to reconstruct randomly masked image patches with the aid of the textual description. In doing so the model is encouraged to learn more expressive representations and textual-visual relations in the latent space. The Attribute Loss, instead, balances the contribution of different types of attributes, defined as adjective-noun chunks of text. This loss ensures that every attribute is taken into consideration in the person retrieval process. Extensive experiments on three commonly used datasets, namely CUHK-PEDES, ICFG-PEDES, and RSTPReid, report performance improvements, with significant gains in the mean Average Precision (mAP) metric w.r.t. the current state of the art. Code will be available at https://github.com/ErgastiAlex/MARS.

Additional Key Words and Phrases: Multi-modal learning, person retrieval, re-identification

1 INTRODUCTION

The integration of text prompts in the re-identification task, called text-based person search (TBPS), has gained lots of interest in the research community lately [1, 16]. In TBPS, textual descriptions are queries used to search *specific identities* in a gallery of images. This is similar yet conceptually different from the standard text-based image retrieval task, in which captions are used to find one or multiple images that best match the given description. Commonly, architectures designed for TBPS include two encoders, one for images and one for the text prompts. The encoders extract a latent code for each modality which can be then aligned using various loss functions such

*Authors contributed equally to this work

Authors' Contact Information: Alex Ergasti, alex.ergasti@unipr.it; Tomaso Fontanini, tomaso.fontanini@unipr.it, Department of Engineering and Architecture, University of Parma, Parma, Italy, IT; Claudio Ferrari, Department of Engineering and Architecture, University of Parma, Parma, Italy, IT, claudio.ferrari2@unipr.it; Massimo Bertozzi, Department of Engineering and Architecture, University of Parma, Parma, Italy, IT, massimo.bertozzi@unipr.it; Andrea Prati, Department of Engineering and Architecture, University of Parma, Parma, Italy, IT, andrea.prati@unipr.it.

ACM 1551-6865/2025/3-ART
https://doi.org/10.1145/3721482

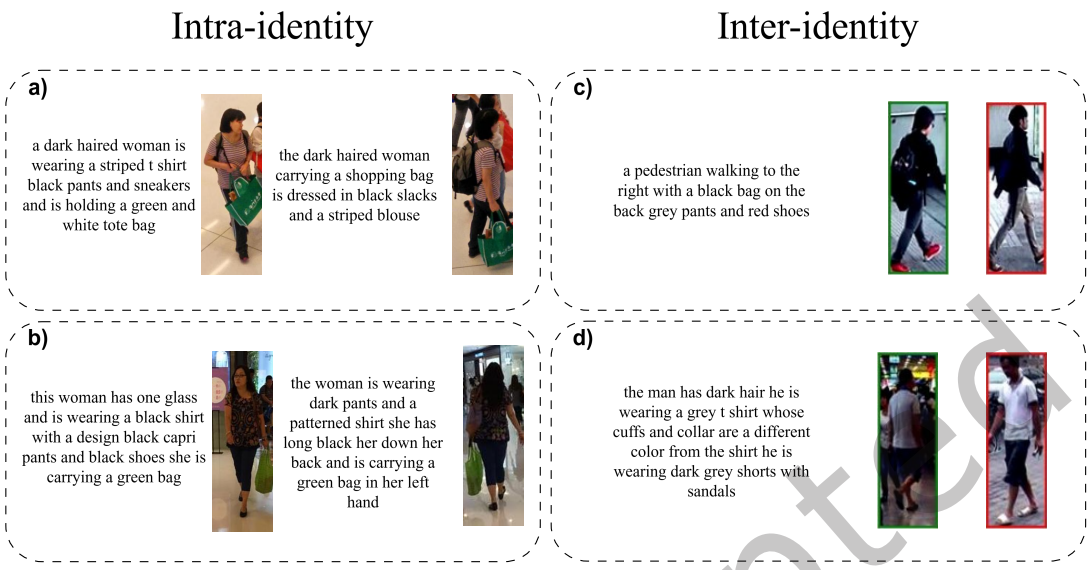

Fig. 1. CUHK-PEDES images and caption. On the left, *a* and *b* are examples of intra-identity variations where the visual attributes of the same person (*e.g.*, pose, illumination, etc..) vary between images. On the right, *c* and *d* are examples of inter-identity variations where a caption can be matched to two identities which look very similar between each others but only one is correct (green for correct match, red for wrong match).

as cross-modal projection matching [29] or contrastive loss [2]. By doing so, textual and visual latent codes are forced to lie in a common space, so that one can use the text embeddings to retrieve the latent code of the most similar image. The most popular choice opted by recent approaches, *e.g.*, [1, 10, 14], is to fine-tune and adapt pre-trained large vision-language models such as CLIP [26], BLIP [11] and ALBEF [12]. This is motivated by the relative small size of datasets commonly used in TBPS, which are typically composed by less than 100k images. The fine grained knowledge that is provided by such large models can be used as a solid starting point to train a TBPS system. Additionally, architectures based on BLIP [14] or ALBEF [1] use a cross-modal encoder that fuses together image and text information via cross-attentions layers and performs an additional matching. More in detail, in such architectures, the searching task is composed of two phases: in the first phase, for each textual embedding, a list of k nearest-neighbor images is obtained; subsequently, a re-ranking of the top k candidates is performed based on the matching results of the cross-modal encoder.

Using text in place of images to perform retrieval opens up both several new possibilities and new challenges. On the one hand, a query image is no longer required, resulting in a more flexible and easy search procedure. On the other hand, text prompts are often vague or ambiguous, and lack the objectivity that images instead can provide. Captions included in standard datasets like “A girl with a black bag and white shirt” lack the necessary unique details that are needed to distinguish similar images. For example, the bag could be both on the right or on the left shoulder, or the shirt could have different details such as logos or textures. Such differences are slight, yet they might correspond to different identities in a given video. This vagueness hinders the quantitative results in a TBPS system in which we care about finding precise identities given the captions and not just the most similar images. We refer to this as *inter-identity* noise (see Fig. 1 on the right). Another key problem is represented by the *intra-identity* variations (see Fig. 1 on the left). The appearance of the same subject in the dataset can vary depending on several factors such as pose, camera position (front or back facing), or illumination.

At the same time, different text descriptions can be used to describe the subject with different level of granularity and ambiguity. These nuisances might have a non-negligible effect; for example, if a person is captured from the back, attributes such as "man/woman" become even more ambiguous.

Several approaches proposed solutions to limit this problem. The most common one consists in building a more fine-grained relationship between image and text embeddings by performing masked language modeling [10, 14]. This is achieved by masking the text prompt and, via cross-attention mechanism, utilizing the image patch embeddings to predict the missing words. Alternatively, RaSa [1] proposed a slightly different solution in addition to masking, which consists in changing some words, and then training the model to recognize which words were changed. In addition to the above, in this work we argue that another problem of current TBPS systems is that the existing text encoding techniques do not fully exploit all the attributes contained in a given text, making the retrieval less precise. Indeed, assigning the same importance to all attributes, especially to the most discriminative ones, is often fundamental to distinguish different identities. In fact, two different subjects, that are yet very similar in appearance, might only be correctly separated by a single small attribute *e.g.*, shoes color, in their description. This is true in particular for long textual description containing several attributes, where we want the TBPS system to balance the contribution of each attribute equally during the retrieval. The attribute loss proposed in this work was designed precisely to push the model to correctly exploit all the attributes.

In this paper, we present a novel TBPS architecture named MARS (Mae-Attribute-Relation-Sensitive) made up of a text encoder, an image encoder, a cross-modal encoder and a masked autoencoder. It introduces two losses during training: firstly, a novel attribute loss is proposed that matches each set of attributes in the captions and the image data. This pushes the cross-modal encoder to consider each attribute in a caption with equal weight and, as a consequence, reduces the uncertainty in the retrieval. Differently from other approaches such as [16], where every word except adverbs, determiners, special characters and numerals is considered an "attribute", we define the attributes in a sentence as a set of words following the structure adjective+noun (*e.g.*, "white shirt"). Each of these set forms an *attribute chunk*. The matching is performed in the output of the cross modal encoders, where the average of the embeddings corresponding to each attribute chunk is classified, strengthening the correlation between textual and image data. Secondly, to further enhance the capability of the text and image encoder, we added a loss inspired by the Masked AutoEncoder (MAE) architecture [7]. In MAE, the input image of the encoder is masked (*i.e.*, some patches are removed) and the decoder is tasked to reconstruct the original images. In MARS the image encoder acts as the MAE encoder, and an additional MAE decoder is used to perform reconstruction. The decoder takes as input also the embeddings extracted from the text to help guide the image reconstruction. By using textual clues, the MAE decoder establishes more robust text-image relationships when reconstructing the missing details in the masked image. In this way, we aim to further enhance the mutual-information encapsulated in both image and text embeddings. Finally, the key contributions of this work are the following:

- **MARS**: a novel TBPS architecture is proposed which is composed by four main components: a text encoder and an image encoder that embed text descriptions and images, a cross-modal encoder with additional cross-attention layers that fuses textual and image embeddings to perform matching and finally, a novel masked autoencoder that reconstructs masked image patches with the help of textual information.
- **Attribute Loss**: We present a novel attribute loss, which aims at improving the matching accuracy between text and image at the attribute level. This loss matches each set of attributes in a given text with the image. This approach provides the model with enhanced capability to equally weigh each attribute in a given description, in order to accurately discriminate between different identities. By doing so, the attribute loss allows the model to put attention on both common and rare attributes in the retrieval process.
- **Masked AutoEncoder Loss**: We present a Masked AutoEncoder loss which aims to reinforce the mutual-information encapsulated in each embedding. It uses the Image Encoder as a MAE encoder, and adds a

new light-weight decoder which accepts the text embedding as additional input to reconstruct the original image.

2 RELATED WORKS

Joining together text and images for the task of text-based image retrieval and tracking was first explored by Shuang, *et al.* [13], who also introduced the CUHK-PEDES dataset. This dataset is composed of a set of pedestrian images paired with a text description which serves as query to retrieve the correct subject. This new dataset and problem to be solved garnered a lot of attention, and several methods were proposed to address it. Zheng *et al.* [30] proposed a novel hierarchical Gumbel attention network to boost cross-modal alignment, while Wang *et al.* [23] introduced a novel multi-granularity embedding learning model. Zhang *et al.* [29] proposed a cross-modal projection matching (CMPM) loss and a cross-modal projection classification (CMPC) loss. Later, Shao *et al.* [20] introduced an end-to-end framework based on transformers to learn, for both text and images, granularity-unified representations. In addition, a set of methods experimented with using auxiliary tasks such as segmentation, pose estimation or attribute prediction to boost the retrieval performance [24, 31]. Wu *et al.* [25] introduced two sub-tasks, image colorization and text completion. The first one helps learning rich text information to colorize gray images, while, in the second one, the model is requested to complete color word vacancies in the captions. Then, Zeng *et al.* [28] proposed a Relation-aware Aggregation Network (RAN) exploiting the relationship between the person and the local objects. Additionally, three auxiliary tasks are introduced: identifying the gender of the pedestrian, discerning the images of the similar pedestrian, and aligning the semantic information between caption and image. Also, a common problem in text-to-image search is the presence of weak positive pairs. This was first tackled by Ding *et al.* [5] who assigned different margins in the triplet loss.

Up until this point, the vision encoder and the text encoder necessary to align the embeddings of the different modalities were trained from scratch. Recently, the use of pretrained vision-language models has caught attention, *e.g.*, in [3, 21, 26]. Cao *et al.* [26] perform an empirical study about using CLIP [18] as backbone for TBPS. Among these, IRRA [10], which was pretrained on CLIP, introduced an Implicit Relation Reasoning module and aims to minimize the KL divergence between distributions of image-text similarity and normalized label matching. Also, IRRA proposed a masked language modelling (MLM) in which a masked set of text embeddings is reconstructed with the aid of image tokens. Additionally, RaSa [1] designed two novel strategies: Relation-Aware learning (RA) and Sensitivity-Aware learning (SA). A concurrent work with RaSa is represented by CADA [14] which focuses of building bidirectional image-text associations. More in detail, it tries to associate text tokens with image patches and image regions with text attributes. The latter is done by modifying the MLM into masking specific attributes and not random words. Finally, RDE [17] recently proposed a novel approach to learn robust text-image associations even with noisy correspondences.

In addition to pretraining on common text-image datasets not specifically tailored to pedestrian identification, Yang *et al.* [27] introduced a novel dataset named MALS (Multi-Attribute and Language Search). The MALS dataset was generated using diffusion models to overcome privacy concerns and annotation costs associated with real-world data collection. To evaluate the effectiveness of this dataset, Yang *et al.* developed a model called APTM (Attribute Prompt Learning and Text Matching Learning). In APTM the authors proposed a new attribute loss, named Image-Attribute Matching (IAM) loss. This loss function is designed to classify image-text pairs (I, T) using concise text descriptions T that contain only partial information about the subject (*e.g.*, "The person wears pants or shorts"). Another approach based on APTM and pretrained on MALS is F-WoRA [22] which introduced a novel filtering algorithm to identify which part of the generated data is crucial for training. On the contrary, in our paper, we propose a structured Attribute Loss with the purpose of pushing the cross-modal encoder to perform the match between image and text using each of the attributes contained in the captions. In particular, our loss does not build a new caption as in [27], but pushes the model to focus more on the attribute embeddings

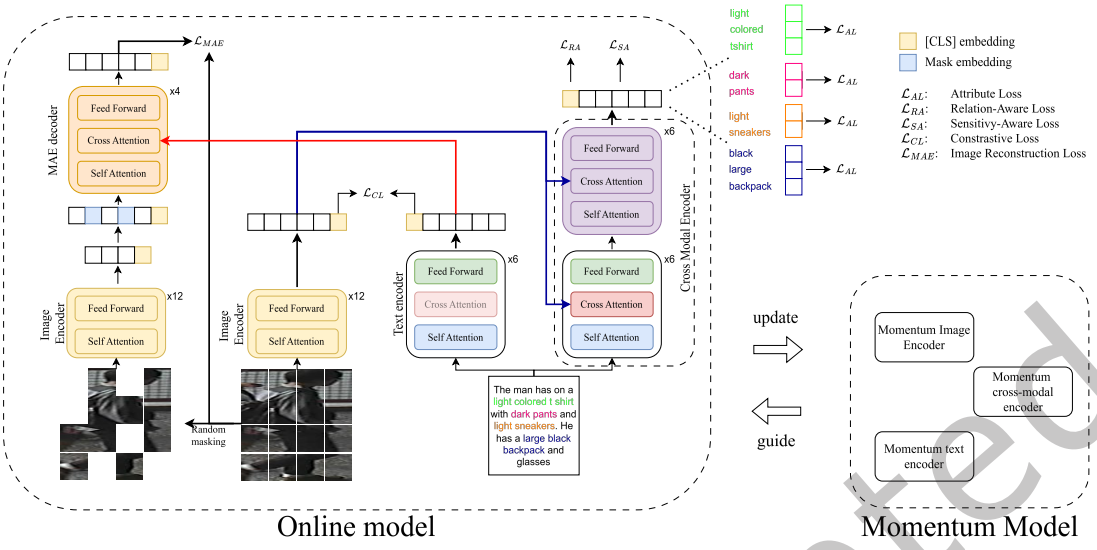

Fig. 2. Overview of the proposed architecture (same color corresponds to shared parameters). Firstly, an input pair of image and text (I, T) is fed to the Image Encoder $\mathcal{E}_v$ and the Text Encoder $\mathcal{E}_t$, respectively, and Contrastive Loss is applied to the obtained embeddings $\mathbf{v}$ and $\mathbf{t}$. Secondly, the MAE Decoder $\mathcal{D}_{mae}$ is trained to reconstruct a sequence of masked image patches into the original unmasked one

. Finally, text is fed to the Cross-Modal Encoder $\mathcal{E}_{cross}$ and the visual embeddings $\mathbf{v}$ are injected into its cross-attention layers. The output of $\mathcal{E}_{cross}$ $\mathbf{f}$ is employed into three different loss functions: (a) the class token f_{cls} is used in the Relation-Aware Loss to learn a matching function between positive and negative image-text pairs, then, (b) given a masked input text T_{mask} Sensitive-Aware Loss is used to identify the masked word and finally, (c) the Attribute Loss is calculated over the embeddings corresponding to attributes chunks in the text.

in an explicit manner. More in details, our model performs an additional matching between image and text based on each of the attributes contained in the sentence.

3 PROPOSED METHOD

In this section, the proposed model architecture, MARS, as well as the training losses are detailed.

3.1 The MARS Architecture

In this paper we propose MARS (Mae-Attribute-Relation-Sensitive), a novel architecture for TBPS. When building the system, we decided to use RaSa [1] as starting point since currently is one of the best TBPS models and we initialized the architecture weights on ALBEF [12].

MARS is composed by four main components (Fig. 2): (a) an Image Encoder $\mathcal{E}_v$ which encodes a sequence of image patches, (b) a Text Encoder $\mathcal{E}_t$ which produces the text embeddings from the captions, (c) a MAE Decoder $\mathcal{D}_{mae}$ which is tasked to reconstruct masked images and, finally, (d) a Cross-Modal Encoder $\mathcal{E}_{cross}$ which computes our proposed attribute loss along with the baseline RaSa [1] losses: Sensitive-Aware and Relation-Aware losses. More in detail, the Image Encoder is a Vision Transformer (ViT) [6] composed by 12 transformer blocks consisting in Self-Attention layers and Feed Forward Layers. The Text Encoder and the Cross-Modal Encoder are based on BERT [4] which is a 12 blocks transformer-based architecture for language understanding. The first 6

blocks of BERT are used as Text Encoder. The Cross-Modal Encoder is composed by all the 12 blocks of BERT, but, differently than previous methods like [1, 12], we equip all its blocks with cross-attention layers instead of only the last 6. By doing so, we can perform the cross-modal encoding using the whole BERT architecture, which helps boosting the matching accuracy as it will be shown in the experiments. Finally, the MAE Decoder is composed by 4 transformer blocks equipped with cross attentions. Additionally, a momentum model is initialized. The momentum model is a slower version of the online model whose weights are obtained using Exponential Moving Average (EMA):

$$\hat{\theta} = m\hat{\theta} + (1 - m)\theta \tag{1}$$

where $\hat{\theta}$ are the weights of the momentum models, while θ are the weights of the online model, and m is the momentum coefficient. This model will be crucial when calculating the losses as explained in Section 3.2.

During training, starting from a image-text pair (I, T), $\mathcal{E}_v$ produces a sequence of image embeddings $\mathbf{v} = \{v_{cls}, v_1, \cdots, v_M\}$ for each of the M image patches, while a tokenized text is fed to $\mathcal{E}_t$ producing a sequence of text embeddings $\mathbf{t} = \{t_{cls}, t_1, \cdots, t_N\}$, being N the number of word. In both $\mathbf{v}$ and $\mathbf{t}$ the first embedding is the class token `[CLS]`. Additionally, a masked version of the image patches of length $L < M$ is embedded using $\mathcal{E}_v$. Then, a set of $K = M - L$ mask embeddings are inserted in the obtained sequence at the masked positions and the whole sequence is fed to $\mathcal{D}_{mae}$ which reconstructs the original image also with the aid of text embeddings $\mathbf{t}$ that are fed in $\mathcal{D}_{mae}$ via cross attention mechanism. Finally, text T is used as input to $\mathcal{E}_{cross}$ while image embeddings $\mathbf{v}$ are injected in $\mathcal{E}_{cross}$ cross attention layers producing the cross-modal embeddings $\mathbf{f} = \{f_{cls}, f_1, \cdots, f_N\}$. The `[CLS]` token of the cross-modal embeddings will be used to perform an additional matching between images and captions.

The evaluation phase is composed of two steps: first, all the image and text embeddings are calculated using the image and text encoder and, for each text embedding, an ordered list of the closest image embedding is obtained by calculating the similarity between the `[CLS]` token of the text and the images. Then, the first k candidates for each text are selected and an additional re-ranking phase is performed considering the matching results of the Cross-Modal Encoder. This additional step allows to further boost the ranking results.

3.2 Baseline Losses

As a baseline training objective for our model, we employ the loss set used in RaSa [1]. Additionaly, our final proposed architecture also introduces two novel losses: an Attribute Loss and a Masked Autoencoder Loss.

Relation-Aware Loss. The Relation-Aware (RA) loss is a modification to the conventional Image-Text Matching (ITM) loss commonly employed in various models [11, 12, 27]. In particular, ITM performs a binary classification between positive and negative image-text pairs. Instead of selecting hard-negative samples at random, the ITM variation, denoted as p-ITM, creates a negative pair set by evaluating embedding similarity and employing this value as the probability of drawing a negative pair. This similarity is quantified using the `[CLS]` token representations from the unimodal encoders (Text and Image Encoder in Fig. 2). The probability of choosing a negative pair is proportional to the similarity of the corresponding image-text `[CLS]` tokens. Consequently, negative pairs exhibiting higher similarity are more likely to be selected, enhancing the robustness of the model in distinguishing between truly-related and unrelated image-text pairs. The loss $\mathcal{L}_{p-ITM}$ is a Cross-Entropy Loss that distinguishes if input pairs (I, T) are positive or negative.

Let $l_c^{itm}(\mathbf{f}_{cls})$ be a fully connected layer applied on the `[CLS]` token of $\mathcal{E}_{cross}(T, \mathcal{E}_v(I))$ which predicts the logit for a given class c. The loss can be calculated as:

$$\mathcal{L}_{p-ITM} = -\frac{1}{3 \cdot N_B} \sum_{(I,T) \in P} \sum_{c \in C} y_c \log \frac{\exp(l_c^{itm}(\mathbf{f}_{cls}))}{\sum_{n \in C} \exp(l_n^{itm}(\mathbf{f}_{cls}))} \tag{2}$$

where C is the set of possible classes, which includes two categories: positive and negative pairs. The variable y_c represents the ground-truth, where $y_c = 1$ if the pair (I, T) belongs to the class c. The set P is built as the union of three subsets, hence the division by 3, each of size N_B, P^{++}, P^{-+}, P^{+-}:

- P^{++} consists of the input batch, where all pairs (I, T) are positive.
- P^{-+} is composed of a negative image I for each text T, sampled randomly with a probability determined by the similarity between t_{cls} obtained from $\mathcal{E}_t(T)$ and v_{cls} obtained from $\mathcal{E}_v(I)$.
- P^{+-} is composed of a negative text T for each image I, sampled randomly with a probability determined by the similarity between v_{cls} obtained from $\mathcal{E}_v(I)$ and t_{cls} obtained from $\mathcal{E}_t(T)$.

Furthermore, the p-ITM loss is expanded by adding a Positive Relation Detection (PRD), formulated as a Cross Entropy Loss, which aims to detect weak positive pairs. During training, the weak positive pairs are built by randomly switching the caption of an image with a caption of a different image having the same identity. Viceversa, we define strong positive pairs as the original pairs coming from the dataset. Let $l_c^{prd}(f_{cls})$ be a fully connected layer applied on the `[CLS]` token of $\mathcal{E}_{cross}(T, \mathcal{E}_v(I))$ which predict the logit for a given class c, then:

$$\mathcal{L}_{prd} = -\frac{1}{N_B} \sum_{(I,T)\in P^{++}} \sum_{c\in C} y_c \log \frac{\exp(l_c^{prd}(f_{cls}))}{\sum_{n\in C} \exp(l_n^{prd}(f_{cls}))} \tag{3}$$

where P^{++} are only positive pairs that can be both weak or strong and C is the number of classes (two in this case), corresponding to strong positive pairs and weak positive pairs. The final RA loss is then computed as:

$$\mathcal{L}_{RA} = \mathcal{L}_{p-ITM} + \lambda_1 \mathcal{L}_{prd} \tag{4}$$

where λ_1 is an hyperparameter used to balance the contribution of $\mathcal{L}_{prd}$.

Sensitive-Aware Loss. Similar to RA loss, Sensitive-Aware (SA) loss is an expansion of the basic Masked Language Modeling (MLM) introduced in [10] that adds a Momentum-based Replace Token Detection (m-RTD). Given a strongly positive pair (I, T), the MLM loss is expressed as a Cross Entropy Loss. Given a masked text T_{mask}, where each word has a probability p of being masked out, the model is trained to predict the correct missing word. Let V represent the set of all possible words in the vocabulary and $l_v(\mathbf{f}_{mask})$ be a fully connected layer applied on each embedding obtained from $\mathcal{E}_{cross}(T_{mask}, \mathcal{E}_v(I))$ which predicts the logit for the vocabulary v. The MLM loss is formulated as:

$$\mathcal{L}_{MLM} = -\frac{1}{N_B} \sum_{(I,T)\in P^{++}} \frac{1}{N_{mask}^t} \sum_{w\in t} m_w \sum_{v\in V} y_v \log \frac{exp(l_v(\mathbf{f}_{mask}))}{\sum_{n\in V} \exp(l_n(\mathbf{f}_{mask}))} \tag{5}$$

where N_B is the batch size, N_{mask}^t is the number of masked words for a given text t, m_w is 1 if the word is masked, otherwise 0 (*i.e.*, $N_{mask}^t = \sum_{w\in t} m_w$) and y_v is a one-hot value on the ground-truth vocabulary. On the other hand, in m-RTD, the focus is on detecting words that have been replaced. To replace the masked word, the momentum model of the MLM is employed, which converges slowly providing less accurate word predictions. The MLM momentum model predicts a word for each masked word, by effectively replacing the masked words with its predictions, and the task of the online model is to identify which of these words have been replaced. The m-RTD loss is based on a Cross-Entropy Loss which teaches the model to distinguish between replaced and non-replaced words. Let C be the set of possible predictions for each word, where a prediction can be either "replaced" or "not replaced", and $l_c(\mathbf{f}_{repl})$) be a fully connected layer applied on each embedding obtained from $\mathcal{E}_{cross}(T_{repl}, \mathcal{E}_v(I))$ which predicts the logit for the class c. The loss function can be expressed as:

$$\mathcal{L}_{m-RTD} = -\frac{1}{N_B} \sum_{(I,T)\in P^{++}} \frac{1}{N_w^t} \sum_{c\in C} y_c \log \frac{\exp(l_c(\mathbf{f}_{repl})))}{\sum_{n\in C} \exp(l_n(\mathbf{f}_{repl})))} \tag{6}$$

where N_B is the batch size, N_w^t is the number of words in a given text t and y_c is the ground-truth. The $\mathcal{L}_{SA}$ is then:

$$\mathcal{L}_{SA} = \mathcal{L}_{MLM} + \lambda_2 \mathcal{L}_{m-RTD} \tag{7}$$

where λ_2 is an hyperparameter used to balance the contribution of $\mathcal{L}_{m-RTD}$.

Contrastive Loss. Contrastive Loss (CL) is the last baseline model loss. As shown by Fig. 2, the contrastive loss is calculated using only the `[CLS]` token of the two encoders, the Image Encoder and the Text Encoder, after passing them into a linear layer to project in a lower dimension space. Given an Image-Text pair (I, T), we obtain v_{cls} from $\mathcal{E}_v(I)$ and t_{cls} from $\mathcal{E}_t(T)$. Then, the two embeddings are fed into the linear layer, obtaining t'_{cls} and v'_{cls}. The same process is replicated also for the momentum model, obtaining $\hat{t'}_{cls}$ and $\hat{v'}_{cls}$. Also, an image queue $\hat{Q}_i$ and a text queue $\hat{Q}_t$ are stored to implicitly enlarge the batch size. The CL is then formulated as:

$$\mathcal{L}_{NCE}(x_1, x_2, Q) = -\frac{1}{|Q|} \sum_{(x,x_+)\in(x_1,x_2)} \log \frac{\exp(s(x, x_+)/\tau)}{\sum_{x_i \in Q} \exp(s(x, x_i)/\tau)} \tag{8}$$

where τ is a learnable temperature parameters, Q is the queue and $s(x, x_+) = \frac{x^T x_+}{||x|| \cdot ||x_+||}$. The image-text constrative loss (ITC) [12, 18] is formulated as:

$$\mathcal{L}_{ITC} = [\mathcal{L}_{NCE}(v'_{cls}, \hat{t'}_{cls}, \hat{Q}_t) + \mathcal{L}_{NCE}(t'_{cls}, \hat{v'}_{cls}, \hat{Q}_v)]/2 \tag{9}$$

Other than $\mathcal{L}_{ITC}$, in RaSa also a intra-modal constrative loss (IMC) is added, which focuses on keeping close the image and text embedding of the same people with respect to the other people.

$$\mathcal{L}_{IMC} = [\mathcal{L}_{NCE}(v'_{cls}, \hat{v'}_{cls}, \hat{Q}_v) + \mathcal{L}_{NCE}(t'_{cls}, \hat{t'}_{cls}, \hat{Q}_t)]/2 \tag{10}$$

The final loss then becomes:

$$\mathcal{L}_{CL} = (\mathcal{L}_{IMC} + \mathcal{L}_{ITC})/2 \tag{11}$$

3.3 Attribute Loss

Our attribute loss, which overview can be seen in Fig. 3, is designed to enhance the model capability to distinguish between matching and non-matching text-image pairs. In particular, we define an attribute in a caption as a chunk of words composed by a noun and its corresponding adjectives (*e.g.*, "white long shirt"). To extract these chunks, SpaCy [9] was employed. The idea behind this loss is that, in captions composed by several attributes, the model is not able to give the right importance to each attributes and potentially could ignore the most discriminative ones. Limiting this effect is crucial since often, due to the vague nature of text description, two people with different identities could be described by very similar texts, differing only for a single attribute. In this case, if most distinctive attributes are neglected, the correct matching between a text description and the correct person could fail, hindering the model accuracy. For this reason, the proposed attribute loss has the objective of limiting these cases, ultimately making the whole system more robust.

In order to do so, given the output of the cross-modal encoder $\mathcal{E}_{cross}$, which takes as input the text T and the image embeddings $\mathbf{v} = \mathcal{E}_v(I)$, for each attribute, *i.e.*, chunk ch of noun-adjective words in a given text T, the average of the corresponding embeddings is calculated as follows:

$$\hat{ch}(T, \mathbf{v}, ch) = \frac{1}{N_w^{ch}} \sum_{w \in ch} \mathcal{E}_{cross}(T, \mathbf{v})_{w^i} \tag{12}$$

where N_w^{ch} is the number of words in a given chunk ch and w^i is the position of the word w in the output of $\mathcal{E}_{cross}$.

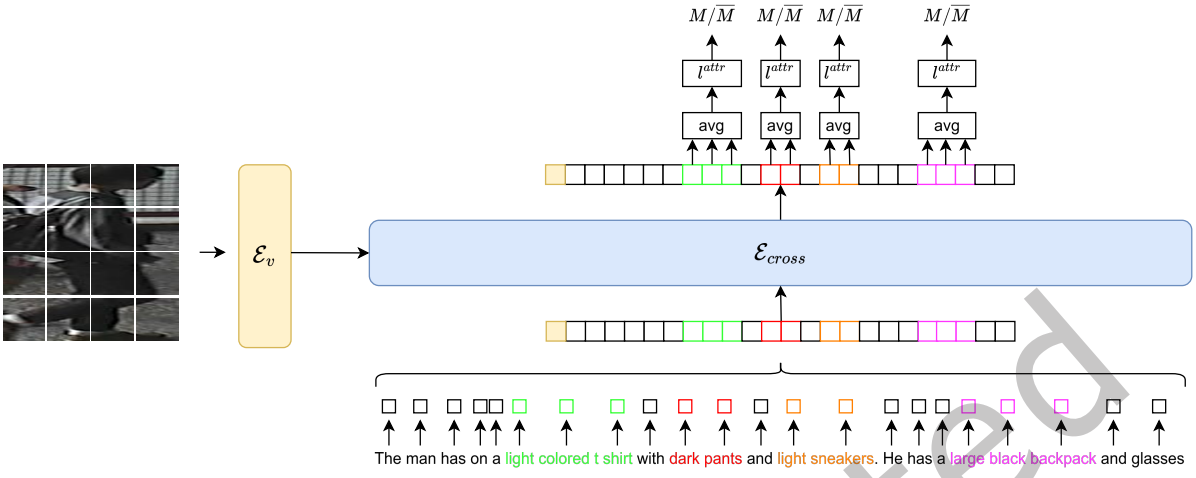

Fig. 3. An overview of the Attribute Loss. Using SpaCy, chunks of sentences containing nouns and related adjectives are identified. Then, after each token is processed by $\mathcal{E}_{cross}$, the average of each chunk embeddings is calculated. For each of them, the model then predicts if the image-chunk pair is a match or not. In the figure, chunks of words with the same color (*i.e.*, green, red, orange and purple) represent the extracted chunks and their corresponding embeddings (each box represents an embedding).

Having this information, is now possible to calculate the proposed Attribute Loss $\mathcal{L}_{\text{AL}}$ for each chunk. More in detail, $\mathcal{L}_{\text{AL}}$ is tasked to perform a matching between each attribute chunk in the caption and the real image. Let N_B be the batch size, N_{ch} the number of chunks in a text T associated with an image I and $l_c^{attr}(\hat{ch}(t, i, ch))$ be the same fully connected layer as the Eq. 2 which predict if the image-text pair (I, T) matches or not. The loss function becomes:

$$\mathcal{L}_{\text{AL}} = \frac{1}{3 \cdot N_B} \sum_{(I,T)\in P} \frac{1}{N_{ch}} \sum_{ch\in t} \sum_{c\in C} y_c \log \frac{\exp(l_c^{attr}(\hat{ch}(T, \mathcal{E}_v(I), ch)))}{\sum_{n\in C} \exp(l_n^{attr}(\hat{ch}(T, \mathcal{E}_v(I), ch)))} \tag{13}$$

Here, C, y_c and P are built as in Eq. 2.

Furthermore, we explored a weighted variant of the loss function. The results of this experiment are presented in Table 2 later in the paper. Specifically, we selected the top 25 most common nouns and adjectives in the CUHK-PEDES corpus (Fig. 4) and calculated the frequency values normalized between 0 and 1. Let α_w denote the frequency of a given word w. If the word is not among the top 25 most common words, we set α_w to 0. We then define the importance weight ω_{ch} for the chunk ch as follows:

$$\omega_{ch} = 1 - \frac{\sum_{w\in ch} \alpha_w}{N_w^{ch}} \tag{14}$$

where N_w^{ch} is the total number of words contained in the chunk. The final weighted attribute loss is formulated as:

$$\mathcal{L}_{\text{weighted-AL}} = \frac{1}{3 \cdot N_B} \sum_{(I,T)\in P} \frac{1}{N_{ch}} \sum_{ch\in t} \sum_{c\in C} \omega_{ch} \cdot y_c \log \frac{\exp(l_c^{attr}(\hat{ch}(T, \mathcal{E}_v(I), ch)))}{\sum_{n\in C} \exp(l_n^{attr}(\hat{ch}(T, \mathcal{E}_v(I), ch)))} \tag{15}$$

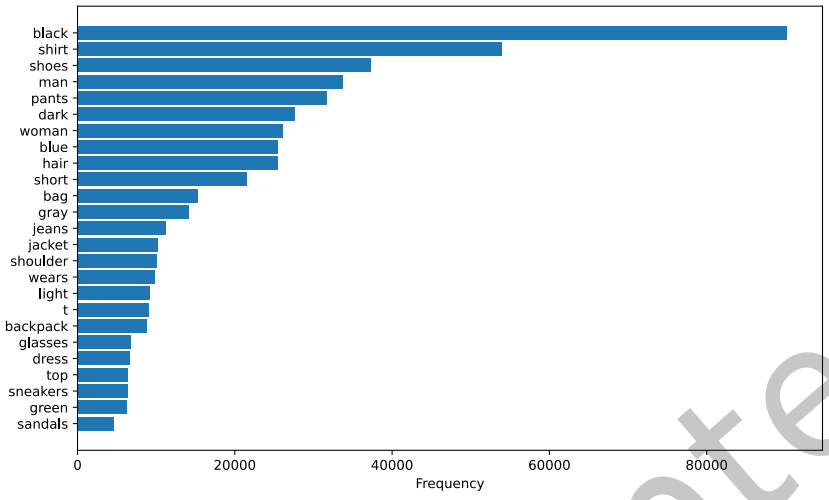

Fig. 4. Top 25 most common nouns and adjectives in CUHK-PEDES computed using SpaCy [9]

As described in Eq. 14, lower importance weights ($\omega_{ch} \rightarrow 0$) are assigned to chunks with very common words and higher importance weights ($\omega_{ch} \rightarrow 1$) are assigned to chunks with uncommon words. This approach is used to downweigh the contribution of very common attributes that match with several different images and therefore identities.

In summary, attribute loss is used to pay attention on the subtle details of a single sentence, improving matching performance using fine-grained details contained in the text that describe an image (*i.e.*, "A pink headset" can be a very uncommon attribute that, if properly considered, improves the model accuracy). As a result, attribute loss helps the model to use the entire given text without losing details. In other words, by distributing the attention evenly, it encourages a more comprehensive understanding of the input data.

3.4 Masked AutoEncoder Loss

Inspired by the masked language model, we have developed a novel loss function based on the Masked AutoEncoder [7] (MAE). MAE was originally used as a self-supervised training technique for transformers. The goal is to reconstruct a sequence of masked image patches back into the original unmasked one. In our case, we customized this technique integrating also text embeddings. More in detail, we inject the text embeddings in the MAE decoder via cross attention layers. The aim is to use the textual information to help the decoder reconstruct the image patches, hence strongly linking together words and visual information.

Given an image-text pair (I, T), we randomly sample patches from the image I with a probability p_{mae} and discard the remaining patches. The selected patches are processed through the Image Encoder $\mathcal{E}_v$ to obtain their corresponding embeddings $\{v_{cls}, v_1, \ldots, v_L\}$, with $L < M$. Prior to feeding these embeddings into the MAE decoder $\mathcal{D}_{mae}$, the embeddings for the removed $K = M - L$ patches are replaced with a learnable mask embedding, thus obtaining a set $\mathbf{v}_{masked} = \{v'_{cls}, v'_1, \ldots, v'_M\}$ of dimension M. The set $\mathbf{v}_{masked}$ is then fed into the MAE decoder $\mathcal{D}_{mae}$, where it is fused with the text embeddings $\{t_{cls}, t_1, \ldots, t_N\} = \mathcal{E}_t(T)$ corresponding to the text T using

cross-attention mechanism to reconstruct the original image. The MAE loss is a reconstruction loss, which is calculated using the mean squared error (MSE) of the removed patches:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N_B}\sum_{i=0}^{N_B}\frac{1}{K}\sum_{j=0}^{M} m_i^j ||x_j^i - \hat{x}_j^i||_2^2 \tag{16}$$

where m_i^j is an indicator variable that equals 1 if the patch was originally removed and thus needs to be reconstructed, and 0 otherwise. Let x_j^i be the original image patch and $\hat{x}_j^i$ be the reconstructed one, then, $\hat{x}_j^i = \mathcal{D}_{mae}(\mathbf{v}_{masked}, \mathcal{E}_t(T))$.

In our case the proposed MAE is trained end-to-end along with all the other components of the model bridging the gap between textual and image information.

3.5 Full Objective and Reranking

Finally, the complete model loss is:

$$\mathcal{L} = \underbrace{\mathcal{L}_{p-ITM} + \lambda_1\mathcal{L}_{prd}}_{\mathcal{L}_{RA}} + \underbrace{\mathcal{L}_{MLM} + \lambda_2\mathcal{L}_{m-RTD}}_{\mathcal{L}_{SA}} + \lambda_3\mathcal{L}_{CL} + \lambda_4\mathcal{L}_{MAE} + \lambda_5\mathcal{L}_{AL} \tag{17}$$

where each λ_* is a weight assigned to a specific loss. During inference, referring to both ALBEF [12] and RaSa [1], considering the high inefficiency of the quadratic interaction operation, we employ a sampling strategy, where we select a subset of k image-text pairs and apply the ITM rank to this reduced set. Specifically, given a text input T, we identify the top-k, with $k = 128$, images by computing the similarity scores $s(t_{cls}, v_{cls})$ and selecting the images with the highest scores. An analysis of how changing this parameter affects both efficiency and accuracy is provided in Section 5.4.

4 EXPERIMENTAL RESULTS

4.1 Experimental Settings

We train our model on a single NVIDIA 4090 GPU for a total of 30 epochs using a batch size of 8. We employ the AdamW optimizer [15] with a weight decay of 0.02. Initial values of the learning rate are set to 10^{-4} for PRD and m-RTD parameters, and 10^{-5} for other parameters. Images are resized to 384×384 (original size is 128×384). Random horizontal flip is applied as augmentation. We set the maximum number of words in BERT to 70. Momentum coefficient m is set to 0.995. The temperature t is set to 0.07, and the queue size utilized in the CL loss is 65536. We set the mask ratio, for the MAE loss, at 75%, thus 75% of image patches are eliminated before going through $\mathcal{E}_v$. We employ the standard BERT [4] for the *MLM* loss, with a masking probability of 15%, while, for the $m-RTD$ loss, a masking probability of 30% is used. Finally, the probability of inputting a weak pair in RA is set to 0.1. We set the λs of the loss described in Eq 17 as $\lambda_1 = 0.5,\ \lambda_2 = 0.5,\ \lambda_3 = 0.5, \lambda_4 = 1,\ \lambda_5 = 2$.

4.2 Metrics

To evaluate our model, we adopt widely-used metrics in TBPS. We evaluate our model with Rank@K, with K=1, 5 and 10. Rank@K counts how many times a model retrieves *at least one* image corresponding to the correct identity in the first K positions. We also calculate the mean Average Precision (mAP), which is computed as follows: let N_T be the number of descriptions in the test set, then we calculate the mAP as the mean average precision for each text t (AP_t):

$$mAP = \frac{1}{N_T}\sum_{t\in T} AP_t \tag{18}$$

AP_t expresses the model accuracy in retrieving the correct images in the early positions for a given text t. It can be calculated as:

$$AP_t = \frac{1}{N_{id,t}} \sum_{k=1}^{N_I} P(k) \cdot \text{rel}(k) \tag{19}$$

where N_I is the number of images in the test set, $N_{id,t}$ is the number of the correct identities for the given text t, $P(k)$ is the precision at the position k, calculated as $\frac{\sum_{i=1}^{k} m_i}{k}$, with $m_i = 1$ if it is a correct match, 0 otherwise and $\text{rel}(k)$ is the indicator function which is 1 if the position k contains a positive match, 0 otherwise.

Compared to ranking metrics, we argue that mAP is most descriptive of the reliability of a retrieval model in real scenarios since it better encapsulates its capability to retrieve more positive matches in top positions. This is especially true for TBPS, where finding all the correct instances corresponding to a specific caption is preferable.

4.3 Datasets

We trained and tested our model on the following three standard datasets used in TBPS literature:

- CUHK-PEDES [13]: composed by 40206 images of pedestrians with 13003 different identities. Each image is paired with 2 text descriptions. The first one contains a coarse description of the image, while the second one is more fine-grained and rich in details. Among all the different identities, 1000 are used for the evaluation phase.
- ICFG-PEDES [5]: includes 54522 pedestrian images of 4102 unique identities. The text information is more fine-grained and identity-centric than CUHK-PEDES. It is divided into a training and a testing set having 34674/19848 images and 3102/1000 identities, respectively.
- RSTPReid [31]: includes 25505 images of 4101 different identities captured with 15 cameras. Each person is represented by 5 images, each having 2 textual descriptions. The dataset is divided in training, validation and testing set having 3701, 200 and 200 identities, respectively.

The three datasets above are all composed of real video-surveillance imagery. Recently, a trend that is explored in the literature is that of considering synthetic data generated via diffusion models for training the vision-language backbones. Given that real datasets are relatively small in size, this option provides a viable solution to increase the size and variability of the training data, without the need of tedious and error-prone manual annotations. For the sake of a comprehensive comparison with recent works, we also employed the following synthetic dataset for pre-training.

- MALS [27]: it is composed by more than 1.5M image-text pairs generated by diffusion models using the annotations of CUHK-PEDES and ICFG-PEDES. After generation, a post-processing phase was performed to remove low quality and noisy images. Finally, the diversity of textual descriptions was increased using BLIP [11] to produce new captions.

4.4 Results Analysis

State-of-the-art works can be broadly divided in two categories: those which use only real datasets and those that instead exploit synthetic data for pre-traininig. To provide a clearer analysis, we separately report the comparisons against such approaches: Table 1.A shows methods that do not employ synthetic data, while Table 1.B includes them.

The first three rows of Table 1.A reports baseline results obtained by directly fine-tuning and applying three pre-trained large vision-language models such as CLIP, BLIP and ALBEF. Then, results of the best state-of-art models are reported. For a fair comparison, they all belong to the family of TBPS models using the aforementioned large-language models. Table 1.B instead presents the comparison between our model and state-of-the-art models based on a novel backbone called APTM, which is pre-trained on the synthetic dataset MALS.

Table 1.A. Results of state-of-the-art models against MARS on CUHK-PEDES, ICFG-PEDES and RSTPReid. All the proposed models in this table are based on backbones pretrained on general datasets.
* model retrained since no checkpoints were available.

		CUHK-PEDES				ICFG-PEDES				RSTPReid			
Model	Backbone	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
-	ALBEF [12]	60.28	79.52	86.34	56.67	34.46	52.32	60.40	19.62	50.10	73.70	82.10	41.73
-	BLIP [11]	64.36	83.36	88.78	58.18	56.16	73.77	80.17	31.59	-	-	-	-
-	CLIP [18]	68.19	86.47	91.47	61.12	56.74	75.72	82.26	31.84	54.05	80.70	88.00	43.41
VGSG [8]	CLIP	67.52	84.37	90.26	-	60.64	76.01	82.01	-	-	-	-	-
CFine [26]	CLIP	69.57	85.93	91.15	-	60.83	76.55	82.42	-	50.55	72.50	81.60	-
TBPS-CLIP [3]	CLIP	73.54	88.19	92.35	65.38	65.05	80.34	85.47	39.83	61.95	83.55	88.75	48.26
RDE [17]	CLIP	75.94	90.63	94.12	67.56	**67.68**	**82.47**	**87.36**	40.06	65.35	83.95	89.90	50.88
CADA* [14]	BLIP	77.20	**90.68**	93.92	68.45	67.38	81.34	85.64	37.81	**67.70**	84.60	89.75	49.95
RaSa [1]	ALBEF	76.51	90.29	94.25	69.38	65.28	80.40	85.12	41.29	66.90	86.50	**91.35**	52.31
MARS (Ours)	ALBEF	**77.62**	90.63	**94.27**	**71.41**	67.60	81.47	85.79	**44.93**	67. 55	**86.65**	**91.35**	**52.92**

Table 1.B. Results of state-of-the-art models against MARS on CUHK-PEDES, ICFG-PEDES and RSTPReid. All the proposed models in this table, except for MARS, are based on the APTM backbone, which is pretrained on the new MALS dataset built for TBPS. These models were then fine-tuned on each proposed dataset. MARS++ refers to the results of our models when pretrained on MALS.

		CUHK-PEDES				ICFG-PEDES				RSTPReid			
Model	Backbone	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
APTM [27]	APTM	76.53	90.04	94.15	66.91	**68.51**	82.99	**87.56**	41.22	67.50	85.70	91.45	52.56
F-WoRA[22]	APTM	76.38	89.72	93.49	67.22	68.35	**83.10**	87.53	42.60	66.85	86.45	91.10	52.49
MARS (Ours)	ALBEF	77.62	90.63	**94.27**	71.41	67.60	81.47	85.79	44.93	67. 55	**86.65**	91.35	52.92
MARS++ (Ours)	ALBEF	**77.89**	**90.68**	93.91	**71.46**	68.50	82.29	86.43	**45.69**	**67.80**	**86.65**	**91.55**	**53.23**

In the subsequent analysis, we will mostly refer to the mAP metric for discussion. We argue that mAP is the metric that most effectively reflects a model capability of accurately retrieving relevant images. Unlike Rank@K, which evaluates if at least one correct image appears within the top k positions, mAP offers a more comprehensive view of the ranking quality across the entire retrieval set. In practical applications where the ground-truth identity in the retrieval set may be unknown, a model that consistently ranks correct images higher across all retrieved items offers substantial benefits. A model scoring an higher mAP identifies a larger set of correct images, and places them in the top positions. In contrast, Rank@K increases even when just one correct image is within the top K, without caring about possible other images of the same individual or the actual rank in the top K. The mAP metric can be said to indirectly measure how robustly a model links a textual description to a specific individual, and thus the robustness of the learned representation. Therefore, the observed improvements in mAP are not only meaningful, but also practically valuable.

Results in Table 1.A show our model reports a consistent improvement over the current state-of-the-art across all proposed datasets, especially in terms of mean Average Precision (mAP). Specifically, on the CUHK-PEDES dataset, our model achieved a mAP of 71.41, which is 2.03 points (2.85%) over the previous best method (69.38). On the ICFG-PEDES dataset, our model obtained a mAP of 44.93, 3.64 points (8.81%) over the second best (41.29).

Similarly, on the RSTPReid dataset, we achieve a mAP of 52.92, a slight increase of 0.61 points (1.15%) over the second best (52.31).

The same considerations apply if we compare MARS with models pretrained on MALS (APTM [27] and F-WoRA [22]), presented in Table 1.B. MARS still outperforms these methods on mAP on each proposed datasets. On CUHK-PEDES, we obtain 4.19 points (6.23%) over of the second best (67.22), on ICFG-PEDES 2.33 points (5.46%) over the second best (42.60), on RSTPReid 0.36 points (0.68%) over the second best (52.92). Nevertheless, we argue that this comparison is unfair since both APTM and F-WoRA employed a large amount of generated image-text pairs during the pretraining phase. For this reason, we experimented using MALS as pre-training dataset for our approach (MARS++ in the table). This led to a further improvement of all the metrics with respect to MARS. This additional experiment further evidences that our architecture trained with the proposed losses is a better alternative to the APTM backbone.

5 ABLATION

In this section, we present a comprehensive analysis of the key components of our model, employing both quantitative metrics and qualitative evaluation to provide a detailed investigation of its performance.

5.1 Quantitative Analysis

Table 2. Ablation study performed on CUHK-PEDES. First two rows represent the baseline (*i.e.*, RaSa) and the baseline trained with caption capped at 70 words instead of 50 words. Other rows (from A1 to A6) show the results of our model with all possible combination of losses (MAE loss $\mathcal{L}_{MAE}$, Attribute Loss $\mathcal{L}_{AL}$ and Full CA that is the text encoder with additional cross attention). Additional ablations with all the losses are provided. The former is the model trained without using the same head for Attribute Loss and Relation-Aware Loss. The latter is the model trained using the rebalanced version of our Attribute Loss.

Model	$\mathcal{L}_{MAE}$	$\mathcal{L}_{AL}$	Full CA	R@1	R@5	R@10	mAP
Baseline (RaSa)	✗	✗	✗	76.51	90.29	<u>94.25</u>	69.38
Baseline with 70 words	✗	✗	✗	77.03	90.24	94.15	70.03
A1	✓	✗	✗	77.08	90.07	94.18	69.88
A2	✗	✓	✗	76.92	90.24	94.01	70.92
A3	✗	✗	✓	77.18	89.90	93.44	70.48
A4	✓	✓	✗	<u>77.63</u>	90.22	94.10	71.35
A5	✓	✗	✓	77.06	90.11	93.92	70.07
A6	✗	✓	✓	77.45	<u>90.48</u>	94.22	**71.46**
Ours w/o shared head	✓	✓	✓	77.18	90.16	93.89	71.20
Ours with AL rebalanced	✓	✓	✓	**77.84**	90.27	94.04	71.19
MARS (Ours)	✓	✓	✓	77.62	**90.63**	**94.27**	<u>71.41</u>

In Table 2 a quantitative comparisons between all the different possible configurations of our model is presented. More in detail, the first two rows report two different RaSa [1] baseline versions with the second one which consider a maximum sentence length of 70 instead of 50 for RaSa. Since the incremented sentence length proved to be better, we have chosen to employ that configuration for all the following experiments. In particular, tests from A1 to A6 represent all the different combinations of training our model with the masked autoencoder loss,

the attribute loss and cross-attention layers in each of the 12 blocks of the Cross Modal Encoder. We decided to comment these results focusing our analysis on the attribute loss and the effect that the other losses have on it.

Surprisingly, this loss alone (test A2) is not able to boost the R@1 score of the model (76.92 vs 77.03), but the mAP is increased (70.92 vs 70.03), meaning that more correct images are found earlier in the retrieval rank. On the contrary, when paired with the masked autoencoder loss (test A4) or the increased cross-attention layers (test A6), the attribute loss is able to improve the overall performance of the model. The motivation for this is twofold. On one side, the MAE loss is able to increase the connection between single words and image patches which benefits also the attribute loss. On the other side, more cross-attention layers means a better interaction between image and text embeddings. Indeed, the importance of the attribute loss is confirmed by the fact that, when the model is trained without it (tests A1, A3 and A5) the quantitative results do not improve w.r.t. the baseline.

Finally, the last two ablations consist in training the attribute loss using a different head than the one used to perform the global matching and re-balancing the attribute loss with weights calculated considering the frequency of the words in the dataset. Quantitative results confirm that sharing the matching head also to perform attribute matching is beneficial to the model and therefore we decided to use this as a final configuration. On the other side, a weighted attribute loss allows to achieve an higher R@1, but it performs worse overall in the other metric and therefore we choose not to use it in our final configuration.

Table 3. Models trained on CUHK-PEDES with one main component removed ($\mathcal{L}_{MAE}$, $\mathcal{L}_{AL}$ or Full CA) tested on ICFG-PEDES

				CUHK → ICFG			
Model	$\mathcal{L}_{MAE}$	$\mathcal{L}_{AL}$	Full CA	Rank@1	Rank@5	Rank@10	mAP
RaSa	✗	✗	✗	50.47	67.37	73.85	24.61
A5	✓	✗	✓	50.62 / +0.15	67.70 / +0.33	74.08 / +0.23	24.64 +0.03
A4	✓	✓	✗	50.98 / +0.51	67.80 / +0.43	74.09 / +0.24	25.14 +0.53
A6	✗	✓	✓	50.88 / +0.41	67.97 / +0.60	74.46 / +0.61	25.49 +0.88
MARS	✓	✓	✓	**51.79 / +1.32**	**68.19 / +0.82**	**74.78 / +0.93**	**25.85 +1.24**

As an additional experiment to further analyze the impact of each component, we perform a cross-dataset test by training three model variants (corresponding to experiment A4, A5, and A6 in Table 2) on CUHK-PEDES and evaluating them on ICFG-PEDES (Table 3). Indeed, when the Attribute Loss is removed, the model performs marginally better w.r.t the baseline (Table 3, model A5). However, pairing the Attribute Loss with either the Full Cross attention layers in the text encoder or the MAE loss resulted in a substantial improvement in the model's performance (Table 3, models A4 and A6). Furthermore, the full model (Table 3, model MARS) has the best overall performance. This shows that the MAE loss and Full Cross Attentions layers serve as auxiliary components, complementing the attribute loss. This experiment proves that each component is necessary to reach optimal performance.

Finally, in Table 4 we performed two additional experiments that demonstrate that the chosen weights for proposed losses are optimal. In both these additional cases the results are inferior to the ones of the final model.

5.2 Efficacy of Attribute Loss

Finally, we want to demonstrate the efficacy of the proposed attribute loss to boost the importance of each attribute chunk in a caption in order to be equally considered w.r.t. the others during the searching process. In order to prove that, we designed the following experiment: firstly, sentences are divided based on the number of

Table 4. Additional tests for the attribute and MAE losses weights in MARS.

Attribute Loss					MAE Loss				
λ_5	R@1	R@5	R@10	mAP	λ_4	R@1	R@5	R@10	mAP
1	77.25	90.28	93.97	71.22	2	77.59	90.59	94.13	71.33
3	77.15	90.29	94.15	71.33	3	77.27	90.44	94.00	71.07
2	**77.62**	**90.63**	**94.27**	**71.41**	1	**77.62**	**90.63**	**94.27**	**71.41**

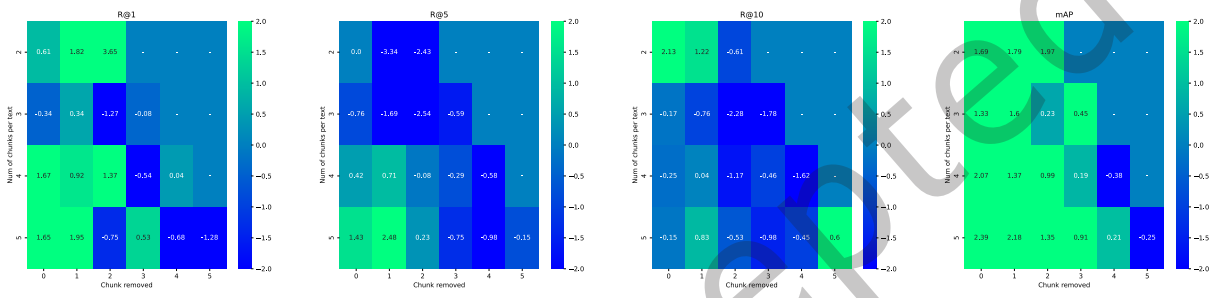

Fig. 5. Results when removing attribute chunks from captions. Each cell contains the difference between the results of our model and the baseline model. Green indicates that our model performs better than the baseline while blue indicates the opposite.

attribute chunks (from 2 to 5); then, different number of chunks (from 0 to 5) are randomly removed from the sentences and the resulting value of R@1, R@5, R@10, and mAP are calculated.

We performed this test both on the baseline model and the proposed model and results can be seen in Fig. 5. More in detail, each cell in the figure represents the difference between our model and the baseline of the corresponding metric value, therefore green colored cells indicate that our model was more resilient than the baseline to the attribute chunk removal, while blue colored cells indicates the opposite.

As expected, thanks to the attribute loss, our model was able to perform much better than the baseline even after the attribute chunk removal. This is especially true for sentences with a higher number of attribute chunks (4 or 5). Indeed, in these cases, removing a single attribute that is crucial in the caption could cause a catastrophic drop in performance in the baseline, while, since in our model each attribute is considered with the same importance, this effect is strongly mitigated. Notably, the mAP metric is basically always better for our model than the baseline. This means that, for each caption, we are always able to retrieve a higher number of the correct identities in better ranking positions. By considering all the attributes equally, some images that were neglected by previous models are now correctly considered during the search.

It is also worth noting that our model performs worse than the baseline especially in cases where a higher number of chunks is removed. In this case, it is important to consider the fact that, when removing a high number of chunks, the accuracy in the retrieval drops dramatically both in the baseline and in our model making it difficult to perform an accurate analysis of the results. At the same time, this indicates that our model performs a search heavily based on attributes and is not able to perform well if no attribute chunk is found in the caption.

Fig. 6. Overview of comparison between top 10 predictions of baseline and our model. Predicted images are ranked from left (*i.e.*, position 1) to the right (*i.e.*, position 10). Our model outperforms the baseline in several pairs, *i.e.*, *a,b,c,d*. In pair *c* it is possible to observe how all predictions are with a bike in it, while this is not true in the baseline. Furthermore, even if in pair *e* our model does not predict the second position correctly, it is easy to observe how a higher mAP is achieve by providing 3 correct matches in top 10 positions compared to 2 correct matches in top 10 of the baseline. Lastly, in pair *f* our model is not able to predict any correct image due to the vagueness of the caption, but is still retrieving images closely related to the text.

5.3 Qualitative Analysis

From a qualitative perspective, in Fig. 6 it is possible to see the difference in ranking between the baseline model (RaSa) and our model. In each image-text pair, the images are the top 10 ordered from left to right, where left is the one with highest matching score. In these examples, a better R@1 and mAP can be seen in image-text pair *a*, where ours model is able to predict the correct image in the first position and also provide another correct image in a higher position compared to the baseline model, the same happens with pair *d*. In pair *b* our model classifies all the first top 5 images correctly. Additionally, in pair *c* our model focuses more on all attributes contained in the given text. Indeed, beside the fact that the first top 4 images are all correct it is possible to observe how also in position 5 and 6 there is a person riding a bike while this behavior is not observable in the baseline. Furthermore, in pair *e* our model is able to achieve higher mAP, even if, compared to the baseline, our second prediction is wrong. Indeed, we are able to predict 3 correct images in the top 10 when the baseline only predicts two correct ones in its top 10. Lastly, we provide in the pair *f* an example of failure of our model, where it is not able to predict any correct image in top 10. However, it is worth noting that all the predicted images of our model are very related to the given caption. In this case this failure is probably due to the intrinsic vagueness of the text captions that often are very difficult to be linked to a specific identity.

Moreover, in Fig 7 several visual comparisons between the baseline model and the model trained with the proposed attribute loss are presented. More in detail, Grad-CAM algorithm [19] was employed to extract attention maps in the cross-modal encoder, each corresponding to the attention of a single word w.r.t the whole person image. In the figure, some attribute chunks were chosen to highlight the effect of the attribute loss. In particular, it can be appreciated how the attention produced using our model is much more consistent over all the words that compose the attribute chuck. This is particularly evident in the “black cross-body bag” and “small wood

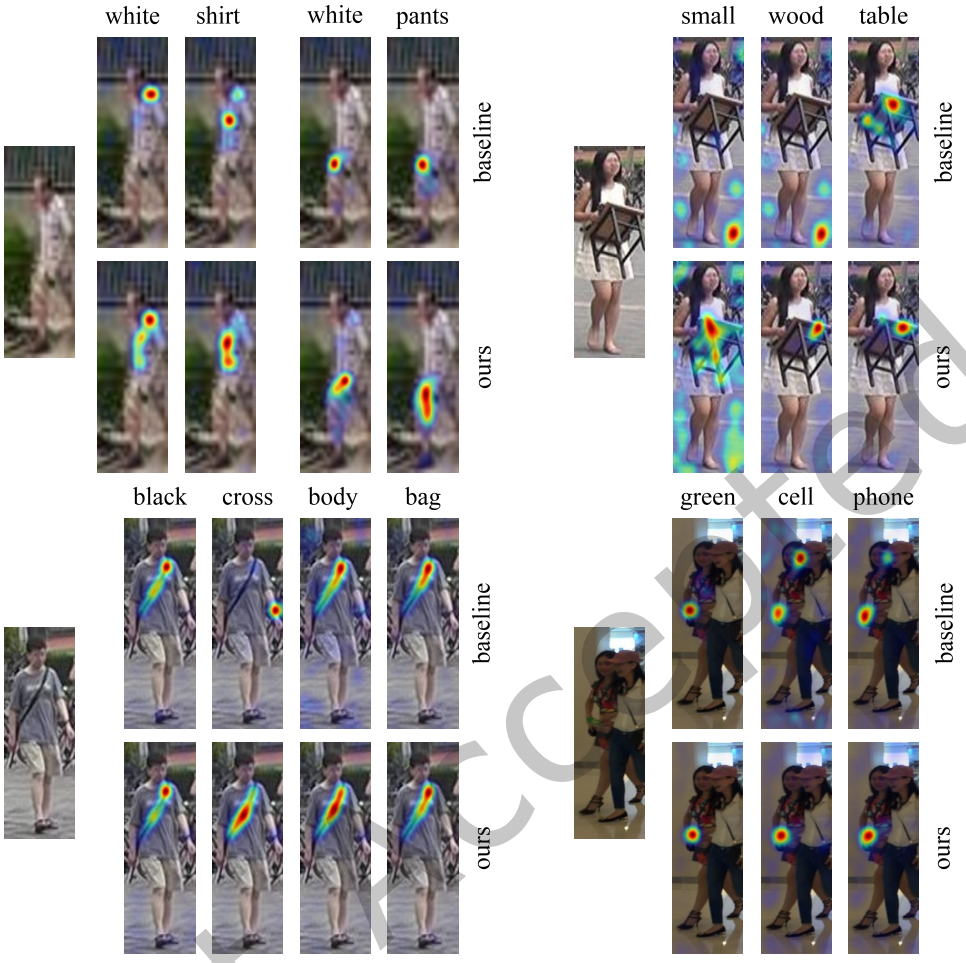

Fig. 7. Visual comparison of cross attention maps generated by the baseline model (top) and our model (bottom) using Grad-CAM [19]. The attention maps illustrate the cross-modal encoder focus on different regions corresponding to individual words in the attribute chunks. The proposed attribute loss leads to more consistent and accurate attention distribution across words.

table" attributes, where the attention is distributed over the correct object for each of the words. In addition, our system allows to generate attention maps that focus much more over the correct attribute. For example, in the attributes "white shirt" and "white pants" the attention maps of our architecture are spread over the corresponding clothes more uniformly than the baseline architecture. On the other side, for the attribute "green cell-phone" no undesired attention noise can be found with our model, whereas the baseline focus is also on random parts of the image. Indeed, these qualitative results help to validate our approach by demonstrating that the proposed training objective helps to precisely link text and image information, which is crucial for the TBPS task.

Finally, in Fig. 8 we present the reconstruction results of the MAE decoder. Each of the input image had 75% of its patches removed. Indeed, even with such a high percentage of masking the proposed module was able to accurately reconstruct the original samples.

Fig. 8. Visualization of MAE decoder results. The mask ratio for each image is 0.75.

5.4 Effect of changing top k for ITM ranking

We select $k = 128$ for ITM ranking by exploring the effect of changing it on several measures, such as time to perform the re-ranking, R@1, R@5, R@10 and mAP. As it can be observed in Fig. 9 increasing k is beneficial for the other metrics, except execution time. In addition, after $k = 128$ the positive effect becomes almost negligible. Indeed, the larger requested time for $k = 256$, which is almost the double w.r.t. $k = 128$, does not justify the accuracy gain, which is only 0.016, considering R@1 only.

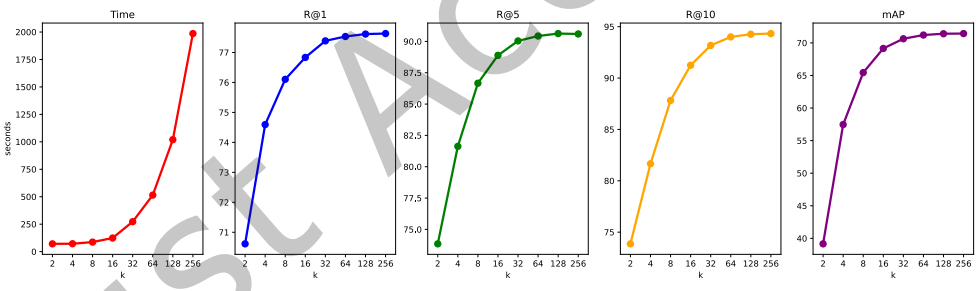

Fig. 9. Impact of varying the sampling parameter k for the ITM ranking on the performance of the model, tested on CUHK-PEDES. The plots show the trade-off between computational efficiency (first plot) expressed in seconds to evaluate the entire test set and the accuracy, expressed as R@1, R@5, R@10 and mAP. We set k equals to 128, since, even if higher values allows the model to obtain better accuracy, these improvements are not justified by the additional required evaluation time.

6 CONCLUSIONS

In this paper we proposed a novel architecture for TBPS named MARS which is composed by a text encoder, an image encoder and a cross-modal encoder, like some of the previous state-of-the-art systems, but, in addition, is also equipped with a masked autoencoder sharing the encoder part with the image encoder and implementing a decoder that takes masked image embeddings as input as well as textual embeddings.

Our proposed MARS architecture brings along a significant improvement in text-based person search. We develop a novel way to address the inter-identity and intra-identity variation, providing a robust solution which is capable to outperform the current state of the art.

Specifically, thanks to the masked autoencoder, we develop a new visual reconstruction loss, which manages to encourage the model to learn a more informative embedding coming from both text and image encoder. Secondly, we equip the whole cross-modal encoder with additional cross attention for the reranking phase. Lastly, we develop a novel attribute loss, which enables the model to focus on every attribute of a given sentence. It is worth noting that, as shown by our ablation this loss alone is not able to push the model to its best, but when coupled with MAE Loss or the new cross model encoder, the attribute loss allows the model to outperform the state of the art.

As a conclusion, all the aforementioned novelties make MARS a model with outstanding performances, especially w.r.t. the mAP. This means that overall, our model is able to rank matching results in earlier positions than previous methods which is crucial in a real world scenario.

As future directions, an interesting option could be delving in 3D reID by reconstructing an approximate 3D shape of the person so to generate virtual novel views to improve the retrieval. Other options include improving the attribute modeling so to obtain a fine-grained separation between positive and negative pairs at the attribute-level, or improving the accuracy in case less detailed, more generic attributes are given.

ACKNOWLEDGMENTS

This research was partially granted by University of Parma through the action Bando di Ateneo 2024 per la ricerca and partially funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16/12/2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU, Project code CN00000023, Concession Decree No. 1033 of 17/06/2022 adopted by the Italian Ministry of University and Research, CUP D93C22000400001, “Sustainable Mobility Center” (CNMS).

Additionally, this work was partially supported by “Partenariato FAIR (Future Artificial Intelligence Research) - PE00000013, CUP J33C22002830006” funded by the European Union - NextGenerationEU through the italian MUR within NRRP.

REFERENCES

[1] 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*. 555–563.

[2] Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. 2023. Text-based Person Search without Parallel Image-Text Data. In *Proceedings of the 31st ACM International Conference on Multimedia*. 757–767.

[3] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. 2024. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 465–473.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[5] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv 2021. *arXiv preprint arXiv:2107.12666* (2021). https://doi.org/10.48550/arXiv.2107.12666

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[8] Shuting He, Hao Luo, Wei Jiang, Xudong Jiang, and Henghui Ding. 2023. VGSG: Vision-Guided Semantic-Group Network for Text-Based Person Search. *IEEE Transactions on Image Processing* 33 (2023), 163–176.

[9] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. (2020). https://doi.org/10.5281/zenodo.1212303
[10] Ding Jiang and Mang Ye. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2787–2797.
[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.
[12] 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.
[13] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1970–1979.
[14] Dixuan Lin, Yixing Peng, Jingke Meng, and Wei-Shi Zheng. 2024. Cross-Modal Adaptive Dual Association for Text-to-Image Person Retrieval. *IEEE Transactions on Multimedia* (2024). https://doi.org/10.1109/TMM.2024.3355644
[15] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7
[16] Kai Niu, Linjiang Huang, Yuzhou Long, Yan Huang, Liang Wang, and Yanning Zhang. 2024. Comprehensive Attribute Prediction Learning for Person Search by Language. *IEEE Transactions on Image Processing* (2024). https://doi.org/10.1109/TIP.2024.3372832
[17] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27197–27206.
[18] 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.
[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
[20] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th acm international conference on multimedia*. 5566–5574.
[21] 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.
[22] Jintao Sun, Zhedong Zheng, and Gangyi Ding. 2024. From Data Deluge to Data Curation: A Filtering-WoRA Paradigm for Efficient Text-based Person Search. *arXiv preprint arXiv:2404.10292* (2024).
[23] Chengji Wang, Zhiming Luo, Yaojin Lin, and Shaozi Li. 2021. Text-based Person Search via Multi-Granularity Embedding Learning.. In *IJCAI*. 1068–1074.
[24] 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: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 402–420.
[25] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. 2021. LapsCore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1624–1633.
[26] 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. https://doi.org/10.1109/TIP.2023.3327924
[27] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4492–4501.
[28] Pengpeng Zeng, Shuaiqi Jing, Jingkuan Song, Kaixuan Fan, Xiangpeng Li, Liansuo We, and Yuan Guo. 2022. Relation-aware aggregation network with auxiliary guidance for text-based person search. *World Wide Web* (2022), 1–18. https://doi.org/10.1007/s11280-021-00953-9
[29] Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*. 686–701.
[30] 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*. 3441–3449.
[31] 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*. 209–217.

Received 30 June 2024; revised 2 December 2024; accepted 17 February 2025