Text Attribute Aggregation and Visual Feature Decomposition for Person Search

Sara Iodice
iodice.sara@gmail.com
Krystian Mikolajczyk
k.mikolajczyk@imperial.ac.uk

Imperial College London
London
UK

Abstract

Person search is the task of retrieving a pedestrian image given a list of text attributes. We investigate a novel mechanism that operates in feature embedding space for matching data across visual and text modalities. We propose a framework (TAVD) with two complementary modules: Text attribute feature aggregation (TA) that aggregates multiple semantic attributes in a bimodal space for globally matching text descriptions with images and Visual feature decomposition (VD) which performs feature embedding for locally matching image regions with text attributes. The results and comparisons to the state of the art on three standard benchmarks demonstrate that our solution is an effective strategy for retrieving person images while retaining the semantic of each query text attribute.

1 Introduction

Research on person re-ID is mostly focused on the problem of matching pedestrian images across different cameras. Several effective solutions have been proposed such as attention [1, 2, 3] and GANs based methods [4, 5, 6, 7], however, these approaches can help only if at least one image of the subject is available, i.e., query image. Unfortunately, this does not occur in many practical scenarios, for example, when reporting a crime, witnesses typically provide only a short textual description of the suspects. To address such cases, re-ID systems should include a more robust matching mechanism that operates across different input modalities, i.e., text and images. There have been several works in the area of person search which consider text input in the form of long and articulate text descriptions [22, 35, 36, 38] or simple lists of text attributes [17, 18]. The second category has the advantage of more data with annotation being available in many popular person or fashion focused benchmarks.
One of the main challenges in a frequent approach that considers person search as a multi-label classification problem, is that multiple labels (attributes) must be predicted for each person category. This leads to over-engineered solutions with a memory complexity growing with the number of training attributes, and limited capacity, as only a fixed number of attributes can be effectively handled. For example, [18] allocates multiple fully connected layers for each attribute, and groups them to handle a large number of attributes. This, however, removes some important semantic differences within each group. In contrast, our

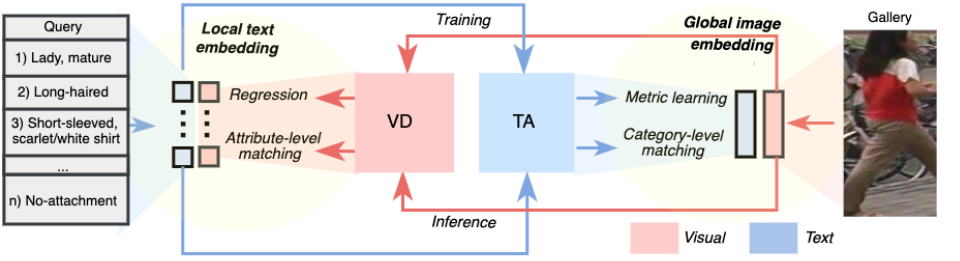

Figure 1: **TAVD**. Text attribute aggregation (TA) and Visual feature decomposition (VD). During training, TA aggregates input text features in a global representation and learns a common metric space; VD decomposes visual features and regresses each component towards the corresponding text embedding. During inference, TA enables global matching between semantic and visual features in a learnt common metric space; VD enables local matching between the input image and the query text attributes. (Best viewed in color).

method preserves the semantic meaning of each single attribute by operating in text embedding space into which visual features are decomposed and locally matched against text attributes. As suggested by [17, 18], another crucial point is to preserve contextual information among the text attributes. We, therefore, propose to also learn a bimodal embedding space where attributes and visual features are holistically represented and matched.

A general overview of our TAVD frameworks is presented in figure 1. TA implements global matching by aggregating input text into a global representation compared to visual features in a common metric space. Concurrently, visual features are decomposed by VD into multiple components and regressed towards the corresponding text embeddings. This enables local matching between image regions and each text attribute. Note that, similar to [18], we consider n attribute categories related to different body parts, but our method can process any number of text attributes by aggregating them in text embedding. Unlike learning visual attribute categories, operating in the text embedding space has the advantage of preserving the semantic differences and relations of attributes. The main contributions of our work are the following: 1) we propose TAVD framework for person search task with novel matching mechanism which operates both globally - leveraging joint representation and context in learnt bimodal space - and locally - comparing image regions to text attributes. 2) in contrast to prior works [17, 18], our approach can handle any number of text attributes while preserving subtle semantic differences between them; 3) to overcome the challenge of limited annotations in standard benchmarks with only a per-person annotation, we propose an augmentation strategy based on synonyms which generates a per-frame annotation; 4) we achieve state-of-the-art results on three person search benchmarks, i.e., Market-1501 [19], DukeMTMC [20] and PA100K [8].

2 Related works

The general goal of text-to-image retrieval is to retrieve an image from a gallery given a text query. The term "person search" was introduced in [18] to define the task of retrieving a person image by text attributes only. Existing approaches for matching image and text can be divided into two main paradigms: category-level and attribute-level methods.

Category-level methods learn a joint feature space where textual descriptions and visual representations can be projected for matching; [28] increases the correlations of positive pairs by minimizing the KL divergence between the normalized true matching probability and the distribution of cross-modal positives; [41] learns to project semantic representation in the visual features space; [43] aligns data from different modalities by minimizing the mean and covariance of corresponding distributions. Inspired by [23, 24, 25] we incorporate a variant of the bi-rank loss [25] we named batch hard bi-triplet loss to enable the matching of both image-to-text and text-to-image and takes advantage of the batch hard triplet configuration [26]. Other methods [29, 30, 31, 32] use adversarial learning to reduce the discrepancy between image and text features. [30, 31, 32] build upon the adversarial framework first proposed by [29]. In [30] textual feature are optimized with the approach from [33] before summarising them with bidirectional LSTM. Interestingly, [31, 32] show the effectiveness of the adversarial strategy for food and recipes matching, although note that adversarial methods may suffer from instability during training.
Example person search methods in this category level group employ adversarial learning to align image and attributes global descriptors [17] or combine the category-level and attribute-level paradigms with a hierarchical approach [18]. Several works [22, 35, 36, 38] have attempted to retrieve images by using natural language description, however, they suffer from the extra challenge of mapping complex sentences with low resolution images. Unlike all these methods, we combine the category-level paradigm with the text embedding paradigm, where we jointly learn a bimodal space for globally matching the semantic text description with visual features, and predict text embeddings for locally matching attributes with image regions.

Attribute-level approaches [10, 11, 18, 34] typically adopt a fully connected layer and a softmax activation to predict the likelihood a certain attribute is present. However, this approach may be inefficient as the number of neurons grows quadratically with the number of attributes and ineffective in case of rare attributes. This has been partly addressed in [9] by learning a disentangled representation for the part and appearance features. Unlike [9] which helps in the case a certain attribute occurs rarely only in certain locations, our method leverages the prior information of the attribute in the natural language model and works with any rare attributes.
Attribute based person search methods [39, 40] focus on matching local attributes and image regions and lack the ability to encode and compare holistic person information.

Vision based person Re-ID methods differ from person search in the final objective which is retrieving a particular ID rather than a person category, i.e., a specific combination of attributes. These methods also benefit from semantic attributes as they provide additional descriptions that are robust against various factors, e.g, illumination condition, pose, and camera view. Specific attributes are particularly useful to deal with ambiguous cases such where persons have a similar appearance but some specific details are different, e.g, shirt logo, hat, and shoe colour. Early works [12, 13] train sequentially the attribute recognition and re-ID tasks by fine-tuning one based on the other. To calibrate the strength of each individual attribute and incorporate dependencies/correlations among them, [14, 15] introduce a re-weighting module which accordingly adjusts the final predictions. Recently, [16, 21] achieved state-of-the-art results on person re-ID and video re-ID benchmarks, respectively. [16] proposes a multitask architecture combining attribute prediction and localiza-

tion; [21] presents an attribute-driven method for feature disentangling and re-weighting based on the attribute recognition confidence.

3 Proposed method

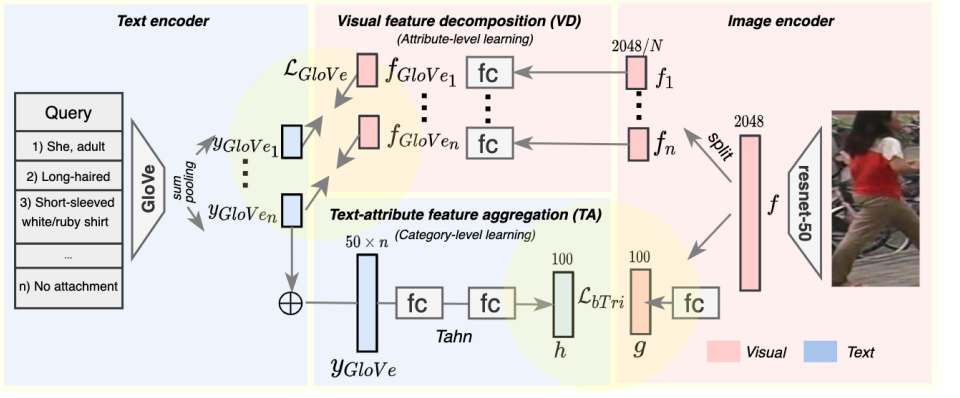

Figure 2: **TAVD framework.** It takes as input a pair formed by a list of text attributes and an image. TA learns a common metric space for comparing global semantic features h and visual features g with the batch hard bi-triplet loss $\mathcal{L}_{bTri}$. VD regresses visual features component $\{f_{GloVe}\}_{i=1}^{n}$ towards corresponding GloVe embeddings $\{y_{GloVe}\}_{i=1}^{n}$. During evaluation time, given the text attribute query, TAVD retrieves the matching image sample from the gallery set evaluating the euclidean distances between the corresponding global and local representations.

In this section, we present our approach for person search using text attributes. Figure 2 shows the processing pipeline. Given a list of text attributes and a gallery image, text attributes encoder aggregates a set of individual attribute embeddings into a global representation. VD implements the opposite flow from image to text and from global to local by decomposing the image representation into local components that describe the text attribute features. The text query is compared to a gallery image in these two spaces focused on local and global representation from both image and text perspective.
We use GloVe model [27] for text embedding that proved successful in many applications with natural language data and ResNet-50 [37] for extracting image features. GloVe model is pre-trained on Wikipedia thus incorporates many synonyms and relations between attributes.

3.1 Text attribute feature aggregation

Retrieving a person image in a gallery by looking for specific local attributes is straightforward for humans. However, rather than simply matching a list of unrelated characteristics, a human also searcher for a global concept emerging from our contextual and prior information regarding those attributes. This motivates our text attribute aggregation module, which computes a holistic representation from local attributes. TA takes as input n attributes and maps them into a semantic $(50-D)$ space with a GloVe model. Each attribute describes a

body part $\{general, head, upperBody, lowerBody, attachment\}$. In case more than one attribute is available for a specific part, the output descriptors are grouped by sum-pooling operation. The n embeddings are then concatenated forming a feature vector of $n \times 50$ components, passed through a first connected layers $50 \times n \times 100$ with a Tahn nonlinearity and finally, a second fully connected layer of 50×50 that maps them into a global $100-D$ text descriptor h. Concurrently, the image is processed by image encoder, which extracts a global $100-D$ image descriptor g thought ResNet-50 backbone and a successive 2048×100 fully connection layer. During training, the proposed framework extracts g and h from a batch of images and text attributes arranged in triplets.

One significant challenge to train the proposed architecture is the lack of annotations in standard benchmarks which provide only a per-person attribute annotation. Typically, there are only a few hundreds of annotated samples and such a small number is insufficient to regress towards $50-D$ text embeddings. Thus to overcome this difficulty, we augment the available annotation with various synonyms as described in section 4.1. To train the TA module we derived the batch hard bi-triplet loss from the bi-rank [25] loss assuming the batch hard triplet configuration [26]:

$$\mathcal{L}_{bTri} = \sum_{g \in G} [m + \overbrace{\max_{h_p} ||g_a - h_p||^2}^{hard\ positive} - \overbrace{\min_{h_n} ||g_a - h_n||^2}^{hard\ negative}]_+ + \sum_{h \in H} [m + \overbrace{\max_{g_p} ||h_a - g_p||^2}^{hard\ positive} - \overbrace{\min_{g_n} ||h_a - g_n||^2}^{hard\ negative}]_+ \quad (1)$$

where g_a, h_p and h_a, g_p are the pairs with the same identity and different modalities; g_a, h_n and h_a, g_n are the pairs with different identities and modalities. Positives and negatives are paired in a batch specifically designed to include C random person classes, with randomly sampled K examples of each class, thus resulting in a batch of CK images. For each anchor in the batch the hardest positive and the hardest negative sample is selected when forming triplets.

3.2 Visual feature decomposition

VD module works concurrently to TA with the image encoder extracting visual features f. These are then split into n components $\{f_i\}_{i=1}^n$ related to different body parts. Each of these components is passed to a separate fully connected layer, which regresses the features towards the corresponding GloVe embeddings of text attributes. This has not only the advantage of preserving the semantic order of the attributes but also provides a bimodal representation which is used during evaluation to retrieve a pedestrian image by a list of text attributes. During the training, we freeze the weights of the GloVe backbone and use the word embeddings y_{GloVe_i} as labels. We train VD with the following loss:

$$\mathcal{L}_{GloVe} = \sum_{i=1}^{n} L_1(f_{GloVe_i}, y_{GloVe_i}) \quad (2)$$

which is the sum of L_1 distances between the regressed embeddings $\{f_{GloVe_i}\}_{i=1}^n$ and the corresponding GloVe labels $\{y_{GloVe_i}\}_{i=1}^n$.

The overall loss is computed for a mini-batch of samples X arranged in triplets:

$$\mathcal{L}(X) = \lambda_{bTri} \cdot \mathcal{L}_{bTri}(X) + \lambda_{GloVe} \cdot \mathcal{L}_{GloVe}(X) \quad (3)$$

Parameters λ_{bTri} and λ_{GloVe} control the contributions from the two modules. During inference, TAVD extracts both a global representation h and local representations $\{y_{GloVe}\}_{i=1}^n$

Market-1501 (attributes)					DukeMTMC (#attributes)					PA100K (#attributes)				
w/o	k=1	k=3	k=5	k=6	w/o	k=1	k=3	k=5	k=6	w/o	k=1	k=3	k=5	k=6
41	73	118	149	161	29	51	85	112	125	30	52	73	82	86

Table 1: **Number of training attributes.** The overall number of binary attributes with and without ($k = 0$) synonyms augmentation in different person search benchmarks. k is the number of synonyms from which we randomly sample for each attribute.

from the text attribute query. It then compares these descriptors with the corresponding visual representations g and $\{f_{GloVe}\}_{i=1}^{n}$ for each image of the gallery set and evaluate the final distance as the sum of euclidean distances between these pairs. The correct match corresponds to the sample having minimum distance from the text attribute query.

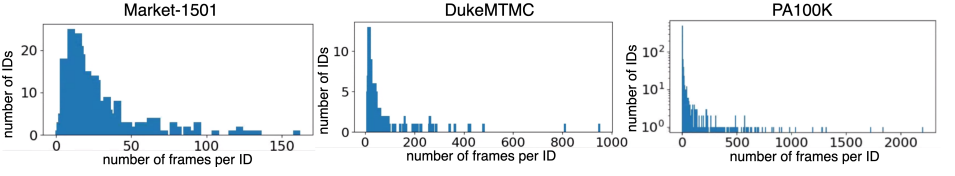

Figure 3: **Frequency of attribute annotations in the datasets**. Note that with ID we refer to a unique set of attributes that annotate a frame. The datasets are highly unbalanced as most IDs have very few image examples, which highlights the importance of data augmentation by adding synonyms to the list of attributes.

4 Experimental results

In this section, we first present the datasets and give more technical details. We then compare our method to the state of the art on several benchmarks and provide a quantitative and qualitative ablation study.

4.1 Datasets and implementation details.

We use three publicly available person search datasets, i.e., Market-1501 [19], DukeMTMC [20], and PA100K [8] following the standard evaluation protocols [17, 18].

Attribute augmentation (synAug). Person search benchmarks provide a person-level attribute annotation with a limited number of text attributes for training in Market-1501 and DukeMTMC. This amount of data is insufficient to train our proposed model. In addition, it is highly unbalanced which we visualize with the histogram of attributes with a given number of image samples in figure 3. We, therefore, propose an augmentation strategy based on synonyms. During training, for each available attribute, we randomly select one of the $k+1$ possible synonyms and use it as an annotation. Consider the attribute "female" as an example. When $k = 5$ we augment it with the synonyms {"maid","women","she","lady","girl"} and during training we annotate each frame containing a female person with a randomly selected synonym from this set. Note that this strategy leads to a frame-level annotation and results in thousands of text attribute annotations that prevent from overfitting. We report the number of binary attributes in table 1 for different values of k. Note that synonyms of an attribute have the same visual representation in images. The full list of synonyms and training

annotations used in our experiments are available online[1].

Implementation details. We used ResNet-50 as image encoder for learning visual features and *GloVe* model pretrained on Wikipedia for learning text attributes features. We adopt $50-d$ GloVe embeddings, which is the smallest GloVe embeddings pre-trained on Wikipedia. We experimented with $100-d$, $300-d$, but performance deteriorates since regressing toward higher-dimensional embeddings causes overfitting. Different dimensionalities of the image descriptor have also been tried, but the best performance is achieved with $100-d$. We set a batch size of $P \cdot M = 64$ with $P = 16$ randomly sampled person IDs and $M = 4$ instances for each person.

Method	Market-1501 r1	r5	r10	mAP	DukeMTMC r1	r5	r10	mAP	PA100K r1	r5	r10	mAP
category-level												
DEM [41]	34.0	48.1	57.5	17.0	22.7	43.9	54.5	12.9	20.8	38.7	44.2	14.8
MMD [42]	34.1	47.9	57.2	18.9	41.7	62.3	68.6	14.2	25.8	38.9	46.2	14.4
DeepCoral [43]	36.5	47.6	55.9	20	46.1	61.0	68.1	17.1	22.0	39.8	48.1	14.1
AAIPR [17]	40.2	49.2	58.6	20.6	46.6	59.6	69.0	15.6	27.3	40.5	49.8	15.2
attribute-level												
GNA-RNN [38]	30.4	38.7	44.4	15.4	34.6	52.7	65.8	14.2	20.3	30.8	38.2	9.3
CMCE [36]	35.0	50.9	56.4	22.8	39.7	56.3	62.7	15.4	25.8	34.9	45.4	13.1
attribute+category												
AIHM [18]	43.3	56.7	64.5	24.3	50.5	65.2	75.3	17.4	31.3	45.1	51.0	17.0
TAVD (ours)	**46.1**	**66.7**	**74.7**	**34.4**	**54.0**	**78.7**	**84.6**	**25.9**	**33.8**	**58.4**	**71.6**	**18.2**

Table 2: Comparison with state-of-the-art methods in Market-1501 [19], DukeMTMC [20], and PA100K [8]. Our approach outperforms all the other methods in terms of rank metrics as well as mAP scores.

We train the network with Adam optimizer, with a learning rate of 0.0003, linear decay to zero over 120 epochs, and $\lambda_{bTri} = 0.9$ $\lambda_{GloVe} = 0.1$. As common practice during the training, we perform image augmentation by using random horizontal flips and re-scaling input images to 256×128 pixels. In addition we adopt synonyms augmentation with $k = 5$ in Market-1501 and $k = 3$ in DukeMTMC and PA100K. Implementation details are also provided in section 3.

4.2 Comparison to the State-of-The-Art Methods

We compare our approach to various solutions for person search, which can be classified into: 1) category-level methods with the goal of learning a joint space for comparing a list of text attributes to images; 2) attribute-level methods focusing on matching of local attributes to image regions. Category-level methods include visual semantic embedding DEM [41], DeepCoral [43], cross-modal matching MMD [42], and GAN based alignment AAIP [17]. Attribute-level methods include GNA-RNN [38] and CMCE [36] which can also handle natural language queries. AIHM [18] combines both strategies to find both global and local correspondences. Our approach extends this idea with a bimodal matching mechanism that jointly operates in a holistic and local space while preserving the semantic relations of the attributes. We achieve state-of-the-art performance in all three benchmarks with $Rank1 = 46.1\%$ in Market-1501 [19], $Rank1 = 54.0\%$ DukeMTMC [20], and $Rank1 = 33.8\%$ PA100K [8]. In the respective benchmarks, we obtain a large margin improvement of $10.1\%, 8.5\%, 1.2\%$ in mAP over the second best AIHM [18]. Note that AIHM has a far more complex hierarchical architecture with multiple branches of fully connected

[1] https://github.com/iodicesara/Text-Attribute-Aggregation-and-Visual-Feature-Decomposition-for-Person-Search

layers for learning semantic and visual features, where the number of branches depends on the number of training attributes/attribute-categories.

4.3 Ablation study

In this analysis, we first demonstrate the importance of learning a bimodal space for matching holistic visual and semantic representations; we then evaluate the effectiveness of the synonyms augmentation strategy (synAug) to cope with the issue of limited attribute annotations in the datasets. We finally demonstrate that learning human attributes in the format of GloVe embedding is a better approach than using a binary representation. We report our evaluation in tables 3 and 4.

			Market1501		DukeMTMC		PA100K	
synAug	$\mathcal{L}_{bTri}$	$\mathcal{L}_{GloVe}$	*r*1	*mAP*	*r*1	*mAP*	*r*1	*mAP*
	✓		0.6	0.3	0.9	1.1	9.2	5.0
✓		✓	0.4	0.5	0.8	1.7	0.0	0.8
✓	✓		45.1	34.4	45.7	25.4	28.7	17.2
✓	✓	✓	**46.1**	**34.4**	**54**	**25.9**	**33.8**	**18.2**

Table 3: Ablation study on Market-1501, DukeMTMC and PA100K. This analysis shows the importance of the loss terms $\mathcal{L}_{bTri}$ and $\mathcal{L}_{GloVe}$ as well as the effectiveness of the augmentation strategy *synAug*. We report performances in terms of *Rank*1($r1$) and *mAP* score.

Person category learning with Bimodal Triplet Loss (TA). Table 3 shows that learning a holistic representation for text attributes in a joint metric space with the image features is essential for person search task. Specifically, removing $\mathcal{L}_{bTri}$ loss leads to performance degradation in all the benchmarks since essential contextual information among the attributes is not exploited. The global loss $\mathcal{L}_{bTri}$ acts as a regulariser and enables the VD to converge.

GloVe embeddings learning (VD). To demonstrate that learning from text attributes in the GloVe space, which incorporates synonyms and semantic relations, is beneficial for person search task we train our TAVD framework with and without $\mathcal{L}_{GloVe}$ loss and compare the final results. When incorporating $\mathcal{L}_{GloVe}$, we note a significant boost of *Rank*1 in all three benchmarks, i.e., 1%, 8.3%, and 5.1% in Market1501 [19], DukeMTMC [20] and PA100K [8], respectively.

Synonyms augmentation. Existing datasets due to a per-person annotation provide a limited number of semantic samples, e.g., only 508 in Market1501, and 300 in DukeMTCM. These are insufficient for TAVD to converge. Instead, the proposed synonym augmentation strategy realises a per-sample annotation and increases the number of semantic samples up to 12,936, 16,522 in Market1501 and DukeMTCM, respectively. Augmenting the data with a number of synonym annotation shows to be essential for our system to learn. In particular, w/o *synAug* the model quickly overfits and performances deteriorate in Market1501 and DukeMTMC. This issue has less impact on large scale dataset PA100K where w/o *synAug* *Rank*1 $= 9.2\%$, however with *synAug* we improve Rank1 to 19.5%. Note that w/o *synAug* all the images of the same person share the same text attribute annotation as well as the same y_{GloVe} descriptor. Thus the hard positive terms in the bimodal triplet loss (equation 1) quickly converge to zero and lead to overfitting. We have observed in our evaluation that it is sufficient to add $k = 1$ synonyms for each attribute to significantly improve the model. We report final evaluation in table 3.

GloVe vs Binary attribute representations. To demonstrate the benefit of learning from

GloVe embeddings we compare the results to a binary representation of text attributes in table 4. *GloVe* is our framework incorporating both $\mathcal{L}_{bTri}$ and $\mathcal{L}_{GloVe}$. *Binary* is the setting with the following modifications: 1) $\mathcal{L}_{GloVe}$ is replaced with the sum of binary cross-entropy loss computed for each attribute; 2) the last fully connected layer followed by a sigmoid nonlinearity outputs a binary vector with a size corresponding to the total number of binary attributes.

We augment the training set annotations and report results in table 4. One can observe that GloVe strategy leads to significantly better results than the Binary representation. In the binary setting, person search is formulated as a multi-category classification problem and it becomes increasingly more challenging when the number of attribute categories to learn grows with limited number of training samples for each attribute. Instead, in "GloVe", we can leverage the prior semantic relations of the text attributes to learn less recurrent attributes in the training set. Without augmentation both learning strategies collapse as in these settings with only few hundreds of semantic inputs. This amount of data is insufficient for the system to learn the bimodal space for globally comparing semantic and visual features, regardless of the type of annotation (GloVe or Binary) used to learn local correspondences between image regions and attributes.

Comparison of VD with attribute-level SOTA methods. Attribute-level methods in Table 2 show superior results as they incorporate other regularisers, e.g., LSTM [38] and co-attention [36]. When $\mathcal{L}_{bTri}$ is included, the gap is reduced. Local features only from VD, which refer to the last row of Table 3 by deactivating TA during evaluation, achieve Rank 1 scores of 12.9,10.2, and 20.1, in Market1501, DukeMTCM, and Pa100K.

	Market-1501				DukeMTMC			
	GloVe		Binary		GloVe		Binary	
	*r*1	*mAP*	*r*1	*mAP*	*r*1	*mAP*	*r*1	*mAP*
w/o synAug	0.6	0.3	0.5	0.4	0.9	1.1	0	1.3
synAug	46.1	34.4	34.6	27.9	54	25.9	32.5	22.9

Table 4: GloVe vs Binary attribute representations. This evaluation compares two training strategies for learning attributes: 1) GloVe (in form of GloVe embeddings); 2) Binary (in form of binary vectors).

Figure 4: Text attribute queries in a),b),c) and d) and their top-9 ranked frame results. We indicate with green/red boxes the correct/false matches and highlight in red text the attributes that are not present in the retrieved frames.

4.4 Qualitative results

To further demonstrate the performance of our method, we visually analyse top ranked results output by our method and show in figure 4. Most of top-9 ranked results match the attributes of the corresponding queries. For example, one can observe that our framework succeeds in retrieving persons with the "hood" in the Rank1 frame (d) where more than half of the image occluded and the “backpack” in the Rank1 frame (a) where only the straps are visible. We found that false retrievals occur: 1) when attributes are not visible due to occlusions or camouflages; 2) due to ambiguity in the semantic meaning of certain attributes; 3) in the case different attributes have similar visual appearance. For example, in examples c and d, one can observe the model can not distinguish subtle differences between violet, blue, and lavender color. An example with a semantic ambiguity is d) where the umbrella may be confused with the hat. This is due to the fact that the training data has no umbrella label.

5 Conclusion

We propose a novel approach (TAVD) which addresses the task of retrieving a pedestrian image from a list of text attributes. We extend the idea of combining category-level and attribute-level paradigms leveraging jointly holistic/contextual information in a learnt bimodal space with the TA module, and the prior information of the attributes in the text embedding space with the VD module.

The proposed method brings the following advantages. 1) It is straightforward as it learns text embedding for each body part in contrast to allocating multiple fully connected layers for each attribute. 2) It is effective as it can handle any number of attributes while preserving the subtle meaning of each attribute. 3) Our TAVD and synonym augmentation overcome the challenge underlined by [18], i.e., learning a rich text embedding from a limited number of semantic samples. 4) In the global space, the batch hard bi-triplet loss with the batch hard triplet configuration is superior to the pairs schema in [18].

We have demonstrated that learning attributes in a local text embedding space better preserves semantic relations than the frequent approach of considering person search as multilabel classification problem, as well as, learning person categories in a global image embedding space is essential for matching holistic visual and semantic representations. Furthermore, we have evaluated the effectiveness of the synonyms augmentation strategy that addresses the issue of limited attribute annotation in person search datasets. We evaluated our proposed methods on three benchmarks (Market-1501, DukeMTMC and PA100K) and achieve an improvement in mAP over the state-of-the-art of up to 10%.

Acknowledgement. This work was supported by UK EPSRC EP/N007743/1 grant.

References

[1] Zheng, M., Karanam, S., Wu, Z. and Radke, R.J., Re-Identification with Consistent Attentive Siamese Networks. In: CVPR (2019)

[2] Li, W., Zhu, X., and Gong, S., Harmonious attention network for person re-identification. In: CVPR (2018)

[3] Xu, J., Zhao, R., Zhu, F., Wang, H., and Ouyang, W., Attention-Aware Compositional Network for Person Re-identification. In: CVPR (2018)

[4] Zhong, Z., Zheng, L., Zheng, Z., Li, S., and Yang, Y., Camera style adaptation for person re-identification. In: CVPR (2018)

[5] Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., and Hu, J., Pose transferrable person re-identification. In: CVPR (2018)

[6] Wei, L., Zhang, S., Gao, W., and Tian, U., Person transfer gan to bridge domain gap for person re-identification. In: CVPR (2018)

[7] Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y. and Kautz, J., Joint discriminative and generative learning for person re-identification. In: CVPR (2019)

[8] Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J. and Wang, X., Hydraplus-net: Attentive deep features for pedestrian analysis. In: ICCV (2017)

[9] Zhao, X., Yang, Y., Zhou, F., Tan, X., Yuan, Y., Bao, Y. and Wu, Y., Recognizing Part Attributes with Insufficient Data In: ICCV (2019)

[10] Gebru, T., Hoffman, J. and Fei-Fei, L.,, Fine-grained recognition in the wild: A multi-task domain adaptation approach. In: ICCV (2017)

[11] Liu, X., Wang, J., Wen, S., Ding, E. and Lin, Y., Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In: AAAI (2017)

[12] Xiao, Q., Cao, K., Chen, H., Peng, F. and Zhang, C., Cross domain knowledge transfer for person re-identification. In: arXiv preprint arXiv:1611.06026 (2016)

[13] Su, Chi and Zhang, Shiliang and Xing, Junliang and Gao, Wen and Tian, Qi, Multi-type attributes driven multi-camera person re-identification. In: Pattern Recognition (2018)

[14] Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C. and Yang, Y., Improving person re-identification by attribute and identity learning. In: Pattern Recognition (2019)

[15] Schumann, A. and Stiefelhagen, R., Person re-identification by deep learning attribute-complementary information. In: CVPR Workshops (2017)

[16] Tay, C.P., Roy, S. and Yap, K.H., AANet: Attribute Attention Network for Person Re-Identifications. In: CVPR (2019)

[17] Yin, Z., Zheng, W.S., Wu, A., Yu, H.X., Wan, H., Guo, X., Huang, F. and Lai, J., Adversarial attribute-image person re-identification. In: arXiv preprint arXiv:1712.01493 (2017)

[18] Dong, Q., Gong, S. and Zhu, X., Person search by text attribute query as zero-shot learning. In: ICCV (2019)

[19] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. and Tian, Q., Scalable Person Re-identification: A Benchmark. In: ICCV (2015)

[20] Ristani, E., Solera, F., Zou, R., Cucchiara, R. and Tomasi, C., Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking In: ECCV (2016)

[21] Zhao, Y., Shen, X., Jin, Z., Lu, H. and Hua, X.S., Attribute-driven feature disentangling and temporal aggregation for video person re-identification In: CVPR (2019)

[22] Yan, F., Kittler, J. and Mikolajczyk, K., Person Re-Identification with Vision and Language In: ICPR (2018)

[23] Jian, Y., Xiao, J., Cao, Y., Khan, A. and Zhu, J., Deep Pairwise Ranking with Multi-label Information for Cross-Modal Retrieval In: ICME (2019)

[24] Wang, L., Li, Y. and Lazebnik, S., Learning deep structure-preserving image-text embeddings In: CVPR (2016)

[25] Liu, Y., Guo, Y., Bakker, E.M. and Lew, M.S., Learning a recurrent residual fusion network for multimodal matching In: ICCV (2017)

[26] Hermans, A., Beyer, L., and Leibe, B., In defense of the triplet loss for person re-identification. In: arXiv preprint arXiv:1703.07737 (2017)

[27] Pennington, J., Socher, R. and Manning, C.D., Glove: Global vectors for word representation In: EMNLP (2014)

[28] Zhang, Y. and Lu, H., Deep cross-modal projection learning for image-text matching In: ECCV (2018)

[29] Wang, B., Yang, Y., Xu, X., Hanjalic, A. and Shen, H.T., Adversarial cross-modal retrieval In: ACM (2017)

[30] Sarafianos, N., Xu, X. and Kakadiaris, I.A., Adversarial Representation Learning for Text-to-Image Matching In: ICCV (2019)

[31] Wang, H., Sahoo, D., Liu, C., Lim, E.P. and Hoi, S.C., Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In: CVPR (2019)

[32] Zhu, B., Ngo, C.W., Chen, J. and Hao, Y., R2GAN: Cross-modal recipe retrieval with generative adversarial network In: CVPR (2019)

[33] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., Bert: Pre-training of deep bidirectional transformers for language understanding In: arXiv preprint arXiv:1810.04805 (2018)

[34] Ji, Z., Sun, Y., Yu, Y., Pang, Y. and Han, J., Attribute-guided network for cross-modal zero-shot hashing In: IEEE transactions on neural networks and learning systems (2019)

[35] Chen, D., Li, H., Liu, X., Shen, Y., Shao, J., Yuan, Z. and Wang, X., Improving deep visual representation for person re-identification by global and local image-language association In: ECCV (2018)

[36] Li, S., Xiao, T., Li, H., Yang, W. and Wang, X., Identity-aware textual-visual matching with latent co-attention. In: ECCV (2017)

[37] He, K., Zhang, X., Ren, S. and Sun, J., Deep residual learning for image recognition In: CVPR (2016)

[38] Li, S., Xiao, T., Li, H., Zhou, B., Yue, D. and Wang, X., Person search with natural language description In: CVPR (2017)

[39] Vaquero, D.A., Feris, R.S., Tran, D., Brown, L., Hampapur, A. and Turk, M., Attribute-based people search in surveillance environments In: WACV (2009)

[40] Layne, R., Hospedales, T.M. and Gong, S., Attributes-based re-identification In: Springer (2014)

[41] Zhang, L., Xiang, T. and Gong, S., Learning a deep embedding model for zero-shot learning In: CVPR (2017)

[42] Tolstikhin, I.O., Sriperumbudur, B.K. and Schölkopf, B., Minimax estimation of maximum mean discrepancy with radial kernels In: NIPS (2016)

[43] Sun, B. and Saenko, K., Deep coral: Correlation alignment for deep domain adaptation In: ECCV (2016)