A Transformer-based Multi-task Model for Attribute-based Person Retrieval

Andreas Specker

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
andreas.specker@kit.edu

Abstract

Person retrieval is a crucial task in video surveillance. While searching for persons-of-interest based on so-called query images gains much interest in the research community, attribute-based approaches are rarely studied. Attribute-based person retrieval takes a person's semantic attributes as input and provides a ranked list of search results that match the description. Typically, such approaches either build on a pedestrian attribute recognition approach or learn a joint feature space between attribute descriptions and image data. In this work, both approaches are combined in a multi-task model to benefit from the advantages of both procedures. Moreover, transformer modules are incorporated to increase performance further. Experimental evaluation proves the effectiveness of the approach and shows that the proposed architecture outperforms the baselines significantly.

1 Introduction

The task of person retrieval aims at finding persons in a large amount of image or video data. It is crucial for effective video surveillance since automatic

person retrieval assists law enforcement agencies in gathering evidence about criminals. Moreover, person retrieval techniques serve as the core component in multi-camera tracking frameworks [20, 13, 25].

Typically, person retrieval algorithms use so-called query images showing the person-of-interest [2, 30, 19] to start a search. However, such images are rarely available in real-world applications since crimes may happen in blind spots of the surveillance cameras. In such cases, attribute descriptions of the criminal gathered from eyewitnesses may be used as a query to start the search. In general, three different procedures are described in the literature. Some approaches directly use natural language queries. These approaches suffer from ambiguity issues of natural language and require complex language processing components. Moreover, merging multiple descriptions is hardly possible. Second, methods learn shared feature spaces between images and textual attribute descriptions. This procedure leads to promising results but discards the semantics of attributes by embedding them into an abstract feature space. Third, pedestrian attribute recognition (PAR) can be leveraged to identify the attributes depicted in the gallery images and match them with the query attributes during retrieval. By that, semantics is preserved, and thus retrieval results are "explainable". In addition, the search for subsets of attributes is enabled without additional computational steps.

This work studies a multi-task model to benefit from both the semantic nature of PAR approaches and the improved performance of shared feature space methods. A model is developed that simultaneously predicts the semantic attributes for images and aligns attributes and corresponding image embeddings. Since recent works indicate that transformer-based models [12, 7] are able to outperform CNN-based equivalents, the model incorporates transformer modules to further enhance performance. During inference, the model benefits from the built-in ensemble since the outputs of the PAR classifier and the joint embeddings are used in combination.

2 Related work

The straightforward approach to attribute-based person retrieval is recognizing the semantic attributes of person images and then comparing the predicted attributes with the query attribute description [27, 21, 15, 22, 24, 23]. This approach provides semantics but is difficult due to the challenging task of recognizing fine-grained, local attributes in low-resolution surveillance imagery.

Further methods align attribute descriptions and image embeddings in a shared cross-modal feature space. In doing so, the large modality gap between attributes and images has to be bridged. Approaches from the literature solve this by using high-dimensional hierarchical embeddings and an additional matching network [5]. Further works aim to match person attributes and images in a joint feature space [31, 1]. Adversarial training is applied to align the different modalities. Jeong et al. argue that this procedure is often unstable and challenging due to the min-max optimization procedure. Thus, they propose an approach that does not employ adversarial training but introduces a modality alignment loss function and a semantic regularization loss to leverage the relations between different attribute combinations explicitly [10].

This work follows a combined approach consisting of an attribute classifier and a joint feature learning method similar to [10].

3 Methods

The structure of the developed multi-task model is depicted in Figure 2.1. It consists of four main parts. The first is the backbone CNN at the bottom, which extracts feature maps from images. Transformer modules further process the outputs. Attribute features and cross-modal embedding features are computed based on the input patches. While the attribute features serve as the input for the PAR classifier, cross-modal image embeddings are aligned to the attribute embeddings produced by the attribute encoder. The fourth part of the model generates the embeddings based on the attribute descriptions that correspond to the input image. In the following, each of the parts and the loss functions are described in detail.

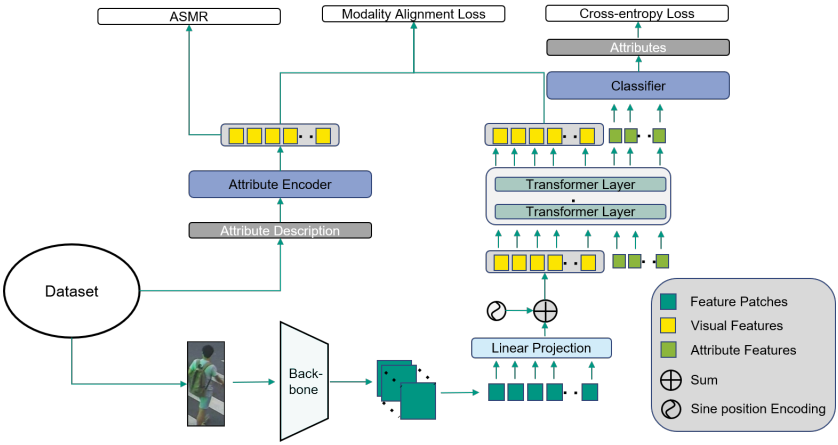

Figure 2.1: Model overview: Structure of the developed multi-task model for attribute-based person retrieval.

Backbone. The backbone model extracts feature maps from the input images. The features maps are then forwarded to the transformer part. Besides the commonly-used CNN backbones such as ResNet-50 [8], transformer-based backbone models gained increasing importance in vision tasks [7, 12]. Transformer backbones may extract better low-level features for small-scale attributes due to their attentive nature and thus improved localization. As a result, experiments in this work are conducted with both types of backbone models. Specifically, the ResNet-50 [8] CNN and the PVTv2-b2 [29] are applied. The latter extracts feature maps more efficiently, although consisting of a similar number of parameters.

Transformer-based Image Encoder. The transformer-based module is incorporated to improve the localization of the different attributes. Visual input tokens are directly extracted from the backbone feature maps. In contrast to the original transformer [28] that works with 1D token sequences as input, 2D visual tokens are leveraged as proposed by Dosovitskiy et al. [6]. For this, feature maps of size $x \in \mathbb{R}^{H \times W \times C}$ are uniformly split into $\frac{H}{P_h} \times \frac{W}{P_w}$ patches. Each patch p

Attribute	Value
# Transformer blocks M	3
# Attention heads	12
Transformer feature dimension d	382
Patch size ($P_w \times P_h$)	2×2

Table 3.1: Transformer architecture: Key facts about the transformer part of the developed model.

of size $P_h \times P_w$. Each patch p is subsequently flattened to obtain a 1D vector with $P_h \cdot P_w \cdot C$ elements. Last, visual tokens v are obtained by projecting flattened patches into a d-dimensional embedding space using a linear function $f : p \longrightarrow v \in \mathbb{R}^d$. Position embeddings [28] pe_i are added to each visual token to encode that each visual token represents a specific area of the input image,

Since the aim is to retrieve cross-modal attribute-image embeddings and features for attribute classification, attribute tokens are additionally incorporated analogous to recent works [18, 9]. For each of the N binary attributes, a learnable, d-dimensional attribute token is concatenated to the visual token. In the following, we refer to these attribute tokens as `[attribute]`. The sequence $T = \{\texttt{[visual]}, \texttt{[attribute]}\}$ then serves as input for the transformer part. It consists of M stacked transformer modules, each of which contains a Multi-head Self-attention bock followed by a Multi-layer Perceptron with layernorm before every block. The outputs are the states of the visual tokens and the N attribute tokens. While the classifier further processes the attribute tokens, global average pooling (GAP) and one fully-connected layer are applied to the visual tokens to obtain the cross-modal 128-dimensional image embedding for cross-modal matching. Details about the parameterization of the transformer blocks are given in Table 3.1.

Pedestrian Attribute Recognition. The states of the attribute tokens `[attribute]` are used as input for the attribute classifier. Fully-connected classification layers generate confidence scores based on the token features of the respective attributes. The result is an N-dimensional vector containing the attributes' confidence scores.

Structure	Size
FC_1	$N \times 512$ ReLU
FC_2	512×128 ReLU
FC_3	128×128

Table 3.2: Structure of the attribute encoder network. It consists of three fully-connected layers and was taken from [10].

Attribute Encoder. The attribute encoder is taken from the work of Jeong et al. [10]. The configuration is provided in Table 3.2. Input is the N-dimensional binarized attribute label vector. Binarization is done by concatenating single elements for binary attributes and one-hot vectors for multi-class attributes. After three fully-connected layers with ReLU activation functions in between, the final cross-modal features with 128 dimensions are obtained.

Loss Functions. Three different loss functions are employed to train the multi-task model: one for the PAR task and two for the modality alignment task.

Most works [14, 16, 17, 26] on PAR consider the task a multi-label classification problem. Analogous to that, multiple binary classifiers with Sigmoid activation functions are used in this work. Therefore, the binary cross-entropy loss function fits the aim of the optimization:

$$L_{PAR} = \frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{N} [y_{i,j} \log(p_{i,j}) + (1 - y_{i,j}) \log(1 - p_{i,j})] * w_j \qquad (3.1)$$

While K denotes the number of training images in the dataset, $y \in \{0, 1\}^N$ stands for the ground truth attribute vector and p for the predicted attribute confidences after the Sigmoid activation layer. Moreover, w_j represents an attribute-specific weighting factor in tackling the problem of imbalanced attribute distributions [14].

Regarding modality alignment (MA), the loss function (L_{MA}) proposed by Jeong et al. [10] that was inspired by the ArcFace [3] loss is utilized. In addition, the Adaptive Semantic Margin Regularizer (ASMR) [10] is applied to balance the distances between attribute embeddings in the joint feature space based on the semantic similarity measured by the Hamming distance between binary attribute vectors. This loss function is termed L_{ASMR} in the following.

The resulting loss function is formulated as follows:

$$L_{total} = L_{PAR} + \lambda_{MA} * L_{MA} + \lambda_{ASMR} * L_{ASMR} \tag{3.2}$$

Further Improvements. Two additional enhancements are investigated in this work and described in the following.

The original calculation of the MA and ASMR losses only considers the attribute sets included in the training data. However, the limited number of training images only constitutes a small excerpt of reality. Since further plausible attribute combinations could be easily generated, further experiments use additional reasonable attribute combinations (AAC) to compute these losses. AAC may improve the generalization concerning persons with previously unseen attribute sets.

The second enhancement relates to the ensemble-like nature of the approach. To support the learning of complementary strengths, the influence of mutual weighting of the two tasks is examined. For example, samples with high loss values in one task are given a higher weight in calculating the loss of the other task. The feedback mechanism should strengthen the complementary aspect of the two tasks by focusing on the weaknesses of the other.

Retrieval. Separate distance matrices are computed based on the PAR outputs and the cross-modal features to perform the retrieval. The Euclidean distance between binary attribute query vectors and confidence scores from the attribute classifier is calculated for PAR. In contrast, query vectors are embedded using the attribute encoder to obtain the distances in the joint feature space. Subsequently, the Cosine distance to the image embeddings of gallery samples is calculated. Last, both distance matrices are normalized by their highest values to achieve scores in the range of 0 to 1, and the weighted sums serve as the final retrieval

Dataset	PETA
# Binary attributes N	35
# Train images K	12,140
# Test images	1,181

Table 4.1: Statistics of the PETA [4] dataset.

distances. The weighting factor λ_{Ret} is applied to the distances from the PAR-based retrieval.

4 Evaluation

In this section, the developed architecture is evaluated to demonstrate its effectiveness and gain insights into the proposed extensions.

Datasets. Evaluation is performed on the PETA [4] dataset. The number of binary attributes N for the dataset is 35. Please note that this follows previous works, which typically use only the 35 most frequent attributes. Table 4.1 provides an overview of the dataset.

Parameters & training setup. Concerning the hyper-parameters of the MA and ASMR loss functions, the values proposed in the original work are applied [10]. PAR baselines are trained as described in [11]. The pre-trained weights of these PAR baselines are used to initialize the backbone. The model is trained using the SGD optimizer with values 0.9 and $5\mathrm{e}{-4}$ for up to 20 epochs. Batches include 16 training samples, and the initial learning rates are set to $1\mathrm{e}{-2}$ for the attribute encoder and $1\mathrm{e}{-3}$ for the remaining model parts. All learning rates are decayed by 0.1 after 10 epochs. Regarding loss weights, λ_{ASMR} values are chosen as proposed in [10], while λ_{MA} is set to 1, i.e., both tasks contribute equally to the total loss. During retrieval, $\lambda_{Ret} = 0.2$ is used to weight the contribution of the PAR task.

Evaluation Metrics. The experiments in this work are evaluated using the Mean Average Precision (mAP) and the Rank 1 score from the Cumulative

Method	mAP	Rank1
PAR Baseline (ResNet-50)	20.0	20.7
PAR Baseline (PVTv2-b2)	21.0	21.5
MA+ASMR (ResNet-50)	20.2	20.0
MA+ASMR (PVTv2-b2)	21.1	20.7
Multi-task Transformer (ResNet-50)	22.7	**22.7**
Multi-task Transformer (PVTv2-b2)	**22.9**	**22.7**

Table 4.2: Comparison of single-task baselines with the introduced multi-task model. Bold numbers denote the best results.

Matching Characteristics (CMC) curve. While the mAP measures the quality of the entire rankings, the Rank 1 accuracy represents the portion of queries in the test set that show a match at the first position in the rank list.

Method	mAP	Rank1
Multi-task Transformer	22.7	22.7
+ AAC	23.1	22.7
+ Feedback MA	23.5	22.6
+ Feedback PAR	**23.7**	23.0
+ Feedback MA&PAR	23.6	23.2
+ Feedback MA&PAR and AAC	**23.7**	**23.3**

Table 4.3: Evaluation of proposed extensions. Bold numbers denote the best results.

Discussion. Table 4.1 contains a comparison of two baseline methods with the proposed approach. Single-task models for PAR [11] and learning a joint feature space [10] serve as baselines. One can observe that the proposed multi-task transformer outperforms the baseline with respect to both evaluation metrics. Using the ResNet-50 as the backbone and the PAR model as a comparison

method, the mAP increases by +2.7% points and the Rank 1 accuracy by +2% points, respectively. Concerning the backbone models, the results indicate no significant difference. While the baseline methods benefit from a transformer-based backbone model, the performance gap is negligible for the proposed approach. This finding indicates that the use of a transformer head is sufficient to achieve good localization of small-scale attributes.

Next, the use of additional attribute categories during training and the cross-task feedback is evaluated in Table 4.3. All the proposed adoptions lead to improvements. In the case of AAC and MA loss feedback, this is limited to the mAP. In contrast, weighting the PAR loss based on the MA loss improves both metrics. Combining both types of feedback further enhances the Rank-1 accuracy and jointly using the additional attribute categories and the feedback leads to the best results.

5 Conclusion and Future Work

In this work, the idea of a transformer-based multi-task model for cross-modal attribute-based person retrieval was presented. Instead of relying on either PAR or learning of joint feature space, both approaches are combined to obtain the flexibility and the semantics of PAR methods while benefiting from the strong retrieval performances of the second procedure. Experimental results indicate that the multi-task model outperforms the single-task baselines.

However, future improvements may include a detailed evaluation of the model architecture and a more sophisticated approach for combining the results of both tasks during retrieval. In addition, it could be observed that optimal training hyper-parameters differ for both tasks. Further investigations are necessary to improve the training procedure and thus the results.

References

[1] Yu-Tong Cao, Jingya Wang, and Dacheng Tao. "Symbiotic Adversarial Learning for Attribute-based Person Search". In: *Proc. European Conference on Computer Vision (ECCV)*. 2020.

[2] Guangyi Chen et al. "Self-Critical Attention Learning for Person Re-Identification". In: *Proc. IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019.

[3] Jiankang Deng et al. "Arcface: Additive angular margin loss for deep face recognition". In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[4] Yubin Deng et al. "Pedestrian attribute recognition at far distance". In: *Proc. ACM Multimedia Conference (ACMMM)*. 2014.

[5] Qi Dong, Shaogang Gong, and Xiatian Zhu. "Person search by text attribute query as zero-shot learning". In: *Proc. IEEE International Conference on Computer Vision (ICCV)*. 2019.

[6] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[7] Kai Han et al. *A Survey on Vision Transformer*. 2021. arXiv: 2012.12556 [cs.CV].

[8] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[9] Shuting He et al. "TransReID: Transformer-Based Object Re-Identification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 15013–15022.

[10] Boseung Jeong, Jicheol Park, and Suha Kwak. "ASMR: Learning Attribute-Based Person Search with Adaptive Semantic Margin Regularizer". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 12016–12025.

[11] Jian Jia et al. “Rethinking of Pedestrian Attribute Recognition: Realistic Datasets with Efficient Method”. In: *arXiv preprint arXiv:2005.11909* (2020).

[12] Salman Khan et al. *Transformers in Vision: A Survey*. 2021. arXiv: 2101.01169 [cs.CV].

[13] Philipp Kohl et al. “The MTA Dataset for Multi-Target Multi-Camera Pedestrian Tracking by Weighted Distance Aggregation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2020.

[14] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. “Multi-attribute Learning for Pedestrian Attribute Recognition in Surveillance Scenarios”. In: *ACPR*. 2015, pp. 111–115.

[15] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. “Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios”. In: *IAPR Asian Conference on Pattern Recognition (ACPR)*. 2015.

[16] Dangwei Li et al. “A richly annotated dataset for pedestrian attribute recognition”. In: *arXiv preprint arXiv:1603.07054* (2016).

[17] Dangwei Li et al. “A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios”. In: *IEEE transactions on image processing* 28.4 (2018), pp. 1575–1590.

[18] Yanjie Li et al. “TokenPose: Learning Keypoint Tokens for Human Pose Estimation”. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

[19] Jicheol Park et al. “Learning Discriminative Part Features Through Attentions For Effective And Scalable Person Search”. In: *IEEE International Conference on Image Processing (ICIP)*. 2020.

[20] Ergys Ristani et al. *Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking*. 2016. arXiv: 1609.01775 [cs.CV].

[21] Walter J Scheirer et al. “Multi-attribute spaces: Calibration for attribute fusion and similarity search”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[22] Arne Schumann, Andreas Specker, and Jürgen Beyerer. “Attribute-based person retrieval and search in video sequences”. In: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2018, pp. 1–6.

[23] Andreas Specker and Jürgen Beyerer. “Improving Attribute-Based Person Retrieval By Using A Calibrated, Weighted, And Distribution-Based Distance Metric”. In: *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2021, pp. 2378–2382.

[24] Andreas Specker, Arne Schumann, and Jürgen Beyerer. “An evaluation of design choices for pedestrian attribute recognition in video”. In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 2331–2335.

[25] Andreas Specker et al. “An Occlusion-Aware Multi-Target Multi-Camera Tracking System”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2021, pp. 4173–4182.

[26] Chufeng Tang et al. “Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale Attribute-Specific Localization”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 4997–5006.

[27] Daniel A Vaquero et al. “Attribute-based people search in surveillance environments”. In: *Proc. Winter Conference on Applications of Computer Vision (WACV)*. 2009.

[28] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[29] Wenhai Wang et al. “Pvtv2: Improved baselines with pyramid vision transformer”. In: *arXiv preprint arXiv:2106.13797* (2021).

[30] Yichao Yan et al. “Learning Context Graph for Person Search”. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[31] Zhou Yin et al. “Adversarial Attribute-Image Person Re-identification”. In: *Proc. International Joint Conferences on Artificial Intelligence (IJCAI)*. 2018.