Pedestrian Attribute Recognition (PAR)

Axel Weissenfeld[1], Bernhard Strobl[1]
[1]AIT Austrian Institute of Technology, Center for Digital Safety & Security
Giefinggasse 4, 1210 Vienna, Austria
{axel.weissenfeld}@ait.ac.at, {bernhard.strobl}@ait.ac.at

Abstract. *Recognising pedestrian attributes is a challenging task for computer vision, particularly when the imaging quality is poor with complex background clutter and uncontrolled viewing conditions. An overview of current popular datasets and there shortcomings is briefly described. The work combines an instant segmentation model with vision transformer, which is trained on a variety of image-text pairs. The approach is promising to enable fast search in large datasets.*

1. Introduction

Pedestrian attributes such as 'blue bag' or 'blond hair' are searchable semantic descriptions to identify persons in image datasets. This kind of search is part of the soft-biometrics field and there are various applications such as face verification, human identification and person re-identification (ReID) [56]. The latter is the process of associating images or videos of the same person taken from different angles and cameras. Pedestrian attributes recognition (PAR) is the task of extracting attributes of given person images, as shown in Fig. 1. These attrubtes can be seen as high-level semantic information which are hopefully more robust to changes in the appearance of people in different images then low-level features, such as HOG [6], LBP [32] or deep features attributes. Although many works have been proposed on this topic, PAR is still an unsolved problem due to challenging factors, such as multi-view camera poses, occlusions, low image resolutions, illumniation challenges, unbalanced data distribution or image blur. Some variations are presented in Fig. 2. In literature PAR is also known as attribute-person recognition (APR) [25] but we use the term PAR.

To address these difficulties, we use deep learning approaches for PAR. In general, deep features obtained by neural networks have stronger discriminative abilities than hand-crafted features. Moreover, we would like to have a neural network, which can be easily adapted to new image sequences and pedestrian attributes. Therefore, in this work we combine an instance-segmentation model with a vision transformer, which was trained on image-text pairs. This combination achieves promising results.

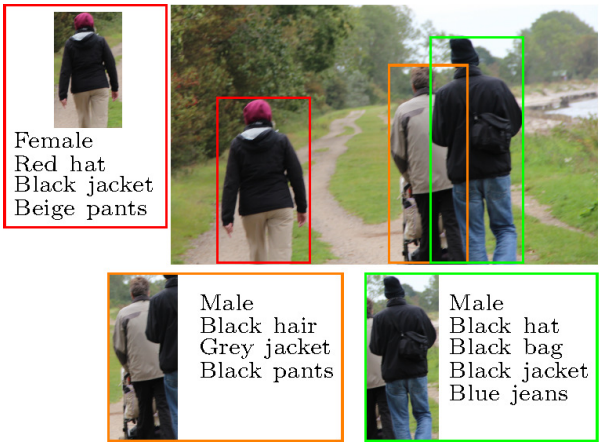

Figure 1. Pedestrian attributes recognition is a key element in video surveillance. The goal is to predict a group of attributes to describe the characteristics of persons. For example, the attributes of the person in the red bounding box are: red hat, black jacket and beige pants.

2. Related Work

The different approaches can be roughly divided into two directions - metric learning and attribute recognition.

2.1. Metric Learning

The objective of metric learning [52, 18] is to learn the metric feature space of persons so that

Figure 2. Example images from the RAPv2 dataset [22]. Top row - from left to right: High quality image, occluded person, blurred image, low resolution image. Bottom row: Various views of the same person.

the distances between similar persons reduce and that of dissimilar persons enlarge. While traditional metric learning algorithms [20] are usually based on linear transformations, recent advances in deep learning provide a powerful tool to learn a task-specific metric. Many metric learning algorithms have been proposed in image retrieval [47], person re-identification [5], face recognition [38, 41, 28, 49], etc. The models directly learns image representations through contrastive loss and triplet loss. Contrastive loss [13] restricts pair inputs and results in distances between similar pairs as close as possible and that of dissimilar pairs to be larger than a threshold. Triplet loss [38] applies triplet as input and ensures the difference between the distance of (anchor, negative) feature and (anchor, positive) feature is larger than a threshold. Beyond triplet loss, quadruplet loss [5] and quintuplet loss [16] are also introduced to improve performance. For face recognition the center loss [48] is applied.

2.2. Attribute Recognition

In PAR, various approaches are available such as global based [1, 21, 8, 39], local parts based [27], visual attention based [30], sequential prediction based [54] methods. Sudowe et al. [39] and Li et al. [21] proposed that jointly training multiple attributes can improve the performance of attribute recognition. The LGNet [27] embeds the attention mechanism [50] in the network, allowing the model to decide where to focus by itself. HydraplusNet [30] proposes an attention based model, which fuses multi-level features to exploit global and local contents of the pedestrian image. ALM [44] localizes attribute-specific regions to achieve better performance by spatial transformer network. Han et al. [14] propose a novel attribute-aware attention model, which can learn local attribute representation and global category representation simultaneously in an end-to-end manner. Localizing by describing [29] is an attribute-guided attention localization scheme where the local region localizers are learned under the guidance of part attribute descriptions. By designing a novel reward strategy, they are able to learn to locate regions that are spatially and semantically distinctive with reinforcement learning algorithm. [37] designs an attention mechanism for aggregating multi-scale features as well as a loss function similar to focal loss [24] in order to tackle the imbalanced data problem. Zao et al. [54] proposes an end-to-end grouping recurrent learning (GRL) model that takes advantage of the intra-group mutual exclusion and inter-group correlation to improve the performance of pedestrian attribute recognition. ReID aims at matching a target person in a set of query pedestrian images. Recent deep learning based ReID approaches achieve promising solutions [46, 10, 4, 2, 11, 34]. Some works combine PAR and ReID information for multi-task learning [40] or assisting the main task [25, 26]. These methods can be briefly divided in two categories: (1) shared backbone and task-independent branches (2) task-independent models and combining high level features in some way (e.g., concatenated FC). For example, Lin et al. [25] use a multi-task network which learns a ReID embedding and at the same time predicts pedestrian attributes, while sharing the same backbone. Sun et al. [40] trains two different branches, the identity one and the attribute one. The identity branch exploits local cues from different regions of the pedestrian body and the attribute branch has an effective attribute predictor.

3. Datasets

The proposed method is evaluated on three publicly available pedestrian attribute datasets and an open image dataset:

Datasets	Resolution	GLVN	BRISQUE
PETA	0.013 (0.008)	20.2	31.5
PA-100K	0.022 (0.022)	24.3	52.7
RAPv2	0.044 (0.028)	28.8	54.5
OpenImages	0.78 (0.74)	32.9	25.5

Table 1. Overview datasets: PETA, PA-100k, RAPv2 and OpenImage. Image resolution calculated in mega pixel. The mediam values of GLVN [36] (normalized gray level variance) and BRISQUE [31] are provided. Note, that a high resolution and a high GLVN value are desired. The lower the BRISQUE value the better the image quality.

- The PETA[1] dataset [8] consists of 19,000 images with 61 binary attributes and 4 multiclass attributes.
- The RAPv2[2] dataset [22] contains 84,928 images which are collected from 25 indoor surveillance cameras, where each image is annotated with 69 fine-grained attributes. Following the official protocol [22], we split the whole dataset into 50,957 training images as well as 16,986 val and 16,986 test images.
- The PA-100K[3] dataset [30] is to-date the largest dataset for pedestrian attribute recognition, which contains 100,000 pedestrian images in total collected from outdoor surveillance cameras. Each image is annotated with 26 commonly used attributes. According to the official setting [30], the whole dataset is randomly split into 80,000 training images, 10,000 validation images and 10,000 test images.
- Open Images[4] is a dataset of ≈9M images annotated with image-level labels, object bounding boxes, object segmentation masks, visual relationships, and localized narratives:

The datasets are briefly analyzed to provide some information about their quality. A summary of some quality quaracteristics are displayed in Tab. 1. Obviously the image resolution and quality of the PAR datasets are rather poor with respect to other commonly used datasets such as Open Images in computer vision.

[1]http://mmlab.ie.cuhk.edu.hk/projects/PETA.html, 11/03/2021

[2]http://www.rapdataset.com/rapv2.html, 11/03/2021

[3]https://github.com/xh-liu/HydraPlus-Net, 11/03/2021

[4]https://storage.googleapis.com/openimages/web/index.html, license: CC BY 2.0

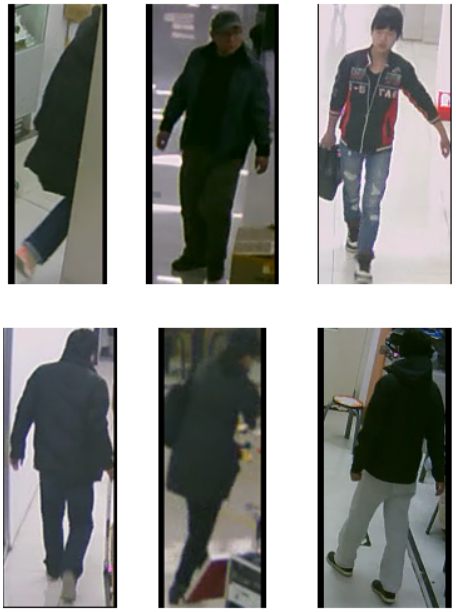

Figure 3. Examples of incorrectly annotated images or unrecognizable annotations of the RAPv2 dataset [22]. Top left: Female, top middle: glasses, top right: hat, down left: attachement box, down middle: glasses, down right: hat.

Novel approaches try to improve the performance by extracting more discriminative features. For this research the openly available popular datasets are used. These datasets, however, must be viewed critically; i.e. there are a large number of identical pedestrian identities in train and test set [17]. This results in a large number of similar images of the same pedestrian identity in the train and test set. Jia et al. [17] analyzed that the proportion of common-identity[5] images is significant in the popular PETA and RAPv2 [22] (updated version of RAPv1 [23]) datasets. They discovered that 57.5% and 31.5% of test set images have similar counterparts of the same pedestrian in the train set of PETA and RAPv2, respectively. Hence, the common-identity issue in the datasets leads to overestimated performance and misleads the evaluation of recent methods [17].

4. Generating novel symthetic data via GANs

Convolutional neural networks (CNNs) have recently become increasingly predominant choices in PAR (as well as ReID) thanks to their strong representation power and the ability to learn invariant deep embeddings. As a result, designing or learn-

[5]Common-identity indicates pedestrian identity exists both in train set and test set. Unique-identity indicates the identity only exists in train set or test set.

Figure 4. Sample images generated based on the dataset Market-1501 [55] and pre-trained GAN model of [57]

ing representations that are robust against intra-class variations has been one of the major targets in PAR. A possibility to enhance robustness of deep learning models is to apply data augmentation during training. Another approach is to collect more data, which can be very tedious and expensive. A rather novel approach is to use Generative adversarial networks (GANs) [12] to generate data from a dataset that is very similar from the original data. With recent progress generative models have become appealing choices to introduce additional augmented data for 'free' [58].

We generated additional data based on work of Zheng et al. [57]. They made their software publicly available[6] for academic research. For generating the images, we used the dataset Market-1501 [55] and their pre-trained model. Same sample images are displayed in Fig. 4. The quality of the generated images is, however, not suffcicient for improving our PAR models. We also tried to finetune the original model to PAR-datasets. But unfortunately the model parameters oscillated and never converged.

5. Evaluation Metrics

This section reviews some common metrics used in the evaluation of PAR methods as described in [23]. In general, two metrices can be calculated at two different levels: label-based and instance-based. The evaluation at label-based considers each attribute independently. The metric adopted in most papers for label-based evaluation is the mean accuracy (mA) [8].

Due to the unbalanced distribution of attributes, mean accuracy (mA) is computed as the average of the classification accuracy of positive and negative examples for each individual attribute i. After that, mA^i is averaged over all attributes as the final recognition rate mA. The evaluation criterion for each attribute i can be formally calculated by:

$$mA^i = \frac{1}{2N}\sum_{j=1}^{N}\left(\frac{TP_j}{P_j} + \frac{TN_j}{N_j}\right) \tag{1}$$

where N is the number of samples; P_j and TP_j are the numbers of positive examples and correctly predicted positive examples, respectively; N_j and TN_j are the numbers of negative examples and correctly predicted negative examples, respectively.

The overall mean accuracy (mA) over L attributes is:

$$mA = \frac{1}{L}\sum_{i=1}^{L} mA^i \tag{2}$$

Instance-based evaluation captures better the consistence of prediction on a given pedestrian image [53], and it includes four metrices: accuracy (Acc), precision (Prec), recall (Rec) rate and F_1 value, as defined below:

$$\begin{aligned}
\text{Acc} &= \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i)|}{|Y_i \cup \hat{Y}_i))|} \\
\text{Prec} &= \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i)|}{|\hat{Y}_i)|} \\
\text{Rec} &= \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap \hat{Y}_i)|}{|Y_i|} \\
F_1 &= \frac{2 * \text{Prec}_i * \text{Rec}_i}{\text{Prec}_i + \text{Rec}_i}
\end{aligned} \tag{3}$$

where N is the number of examples, Y_i is the ground truth positive label of the i'th example, $\hat{Y}_i$ returns the predicted positive labels for i'th example and $|\cdot|$ means the set cardinality.

6. Methodology and evaluation

Many models proposed in literature are optimized exactly for particular datasets. And quite often a sep-

[6] https://github.com/NVlabs/DG-Net, 11/03/2021

arate model is trained for each dataset, which makes it hard for real-world applications. Therefore, we wanted to use models that generalize well and can be quickly applied to new datasets.

Training large networks by scratch is really expensive nowadays. Some examples shall illustrate the effort needed for training large models. For instance, XLNet1 [51], a model for language understanding, was trained on 512 TPU v3 chips for 5.5 days. The total costs are approximately over $500,000 (= 512 * 5.5 days * 8$/hour[7]). Noisy student EfficientNet-L22 was trained with 300MM images on 256 TPU v3 chips for 6 days, which costs ≈$295,000. The largest CLIP model [35] (RN50x64) was trained for 18 days on 592 V100 GPUs. The price of a single V100 GPUs is €10,500[8].

6.1. DeepFace

Deepface[9] is a lightweight face recognition and facial attribute analysis (age, gender, emotion and race) framework for python. It is a hybrid face recognition framework wrapping state-of-the-art models: VGG-Face [33], Google FaceNet [38], OpenFace [3], Facebook DeepFace [43], DeepID [42], ArcFace [7] and Dlib [19]. Those models already reached and passed the human level accuracy. The library is mainly based on TensorFlow and Keras. Some results are displaye in Tab. 2. Note, that faces in PAR datasets are really small, so it is rather challenging to extract semantic information based only on the face.

6.2. ViT

The vision transformer (ViT) introduced by Dosovitskiy et al. [9] is an architecture directly inherited from Natural Language Processing [45], but applied to image classification with raw image patches as input. Their paper presented excellent results with transformers trained with a large private labelled image dataset containing 300 millions images. The paper concluded that vision transformers do not generalize well when trained on insufficient amounts of data. The training of these models involved extensive computing resources.

We used a Keras implementation[10] of the ViT model, which is distributed under the Apache License 2.0[11]. For the experiments, we used the pre-trained vit_b32 model with over 87 million parameters, which we fine-tuned.

Some classification results are presented in Tab. 3. In particular common attributes such as 'female' or 'glasses' are precisely recognized.

6.3. CLIP

We decided to use a novel network [35], denoted as CLIP (Contrastive Language-Image Pre-Training), that has been trained on over 400 million text image data from the Internet and also performs very well on new datasets without any fine-tuning. This network allows a simple textual search in the image data by entering keywords (sentences). The possibility of extensive textual input is especially beneficial for PAR analysis, since different attributes can be combined. For example, a man with a black backpack can be searched for by entering "A photo of a man with a black backpack.". The network returns a confidence score (probability) that an image has the searched attributes.

The framework as well as the models are published unter the MIT license[12] and can be used for commercial use. We used the ViT B/16 model, which has 149 million parameters. We analyzed the CLIP model with the PAD datasets (Tab. 4) and the model achieves for common attributes impressive results. Since the model was trained on text-image pairs retrieved from the internet, the model achieves very good results with ordinary image datasets. Note that zero-shot CLIP is quite weak on several specialized, complex, or abstract tasks such as satellite image classification (EuroSAT and RESISC45) or lymph node tumor detection (PatchCamelyon). On the other hand does the model provide the opportunity to search for rare attributes such as gas mask (Fig. 6) or protective suit (Fig. 7) .

The finally developed PAD method combines two models. The first model ensures, that the people in the image are segmented (instance segmentation). Instance segmentation is a combination of object detection and segmentation. While object detection identifies objects in the image data, segmentation assigns an object class to each pixel. For the instance segmentation used here, a network architecture called Mask R-CNN was used [15]. The Mask R-CNN model was fine-tuned on the OpenImages[13]

[7] https://cloud.google.com/tpu/pricing, 19.04.2021
[8] https://geizhals.at/, 19.04.2021
[9] https://github.com/serengil/deepface, 11/03/2021
[10] https://github.com/faustomorales/vit-keras, 11/03/2021
[11] https://www.apache.org/licenses/LICENSE-2.0
[12] https://opensource.org/licenses/MIT
[13] https://storage.googleapis.com/openimages/web/index.html

Attr.	Samples	Prec.	Rec.	F_1	mA
Gender	45336	0.85	0.17	0.28	0.75
Age <16	753	0.0	0.0	0.0	0.99
Age 17-30	34178	0.43	0.55	0.48	0.54
Age 31-45	46590	0.60	0.53	0.56	0.54
Age 46-60	2968	0.2	0.0	0.0	0.96
Age >60	192	0.0	0.0	0.0	0.997

Table 2. RAPv2 - Deepface [43] results.

	PA100k					RAPv2				
Attr.	Samples	Prec.	Rec.	F_1	mA	Samples	Prec.	Rec.	F_1	mA
Female	45336	0.77	0.77	0.77	0.82	26535	0.88	0.79	0.83 (0.02)	0.90 (0.01)
Hat	4206	0.82	0.30	0.43	0.62	1324	0.70	0.64	0.68 (0.05)	0.68 (0.03)
Glasses	18662	0.70	0.72	0.71	0.71	5700	0.63	0.70	0.65 (0.06)	0.64 (0.02)
Bags	49980	0.70	0.58	0.63	0.68	16482	0.69	0.71	0.68 (0.06)	0.68 (0.04)

Table 3. Evaluation results on four attributes of the fine-tuned ViT model on the PA-100k and RAPv2 datasets.

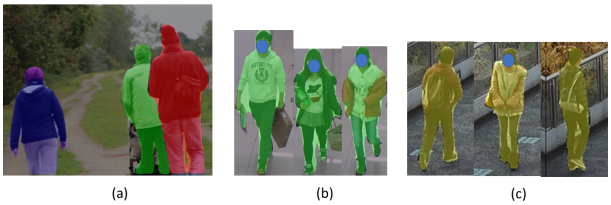

Figure 5. Examples of Mask R-CNN output. (a) Even occluded individuals can be identified very well and segmented with pixel precision. (b) Results from the RAPv2 [22] and PETA [8] datasets: The segmentation even works on small images with high accuracy and reliability.

and Coco[14] dataset. The instance segmentation allows only the image portion containing a person to be further processed. The background, which may disturb the classification, is eliminated and higher image resolutions can be processed, which also improves the classification of the second model. Some classification results of the CLIP model are presented in Fig. 6 and 7.

7. Conclusion

This paper proposes a novel PAD methods to extract useful semantic information for real-world applications. The main idea is to combine an instance segmentation model with a vision transformer, which connnects images and texts. The approach achieves promising results.

[14]https://cocodataset.org/#home

Acknowledgment

This research was funded by the City of Vienna through the Wirtschaftsagentur Wien.

References

[1] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia. Multi-task cnn model for attribute prediction. *IEEE Transactions on Multimedia*, 17(11):1949–1959, 2015. 2

[2] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3908–3916, 2015. 2

[3] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016. 5

[4] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8351–8361, 2019. 2

[5] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017. 2

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 1

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference*

	PA100k					RAPv2				
Attr.	**Samples**	**Prec.**	**Rec.**	F_1	**mA**	**Samples**	**Prec.**	**Rec.**	F_1	**mA**
Female	45336	0.89	0.93	0.91	0.90	26535	0.95	0.98	0.96	0.95
Hat	4206	0.05	0.79	0.08	0.31	1324	0.03	0.72	0.05	0.57
Glasses	18662	0.47	0.30	0.37	0.81	5700	0.76	0.18	0.29	0.94
Bags	49980	0.19	0.66	0.30	0.50	16482	0.26	0.64	0.37	0.60

Table 4. Zero-shot evaluation of PA-100k and RAPv2 datasets with CLIP.

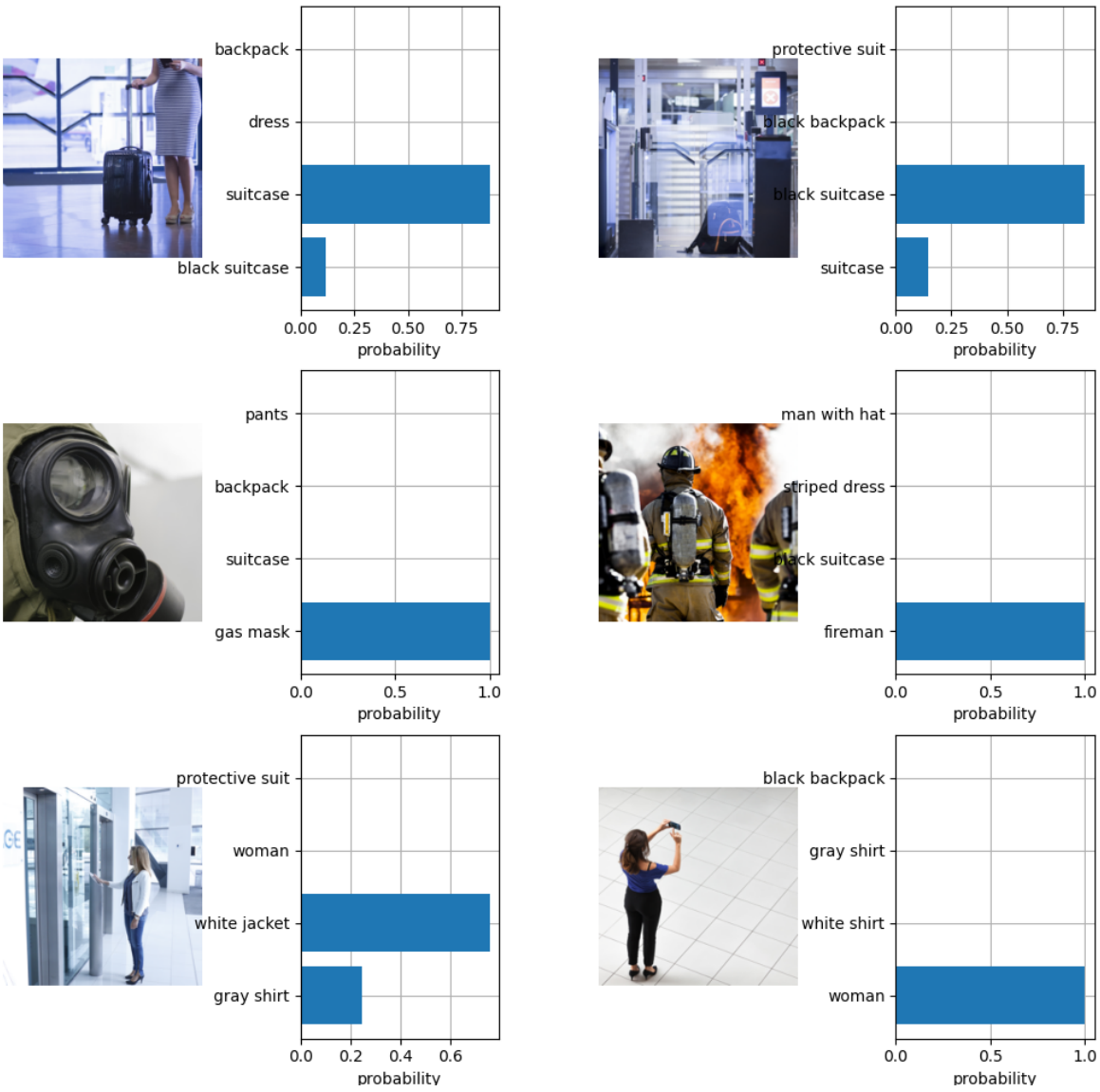

Figure 6. Sample results of CLIP model. Most likely attribute has highest probability.

on Computer Vision and Pattern Recognition, pages 4690–4699, 2019. 5

[8] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792, 2014. 2, 4, 5

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[10] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol-

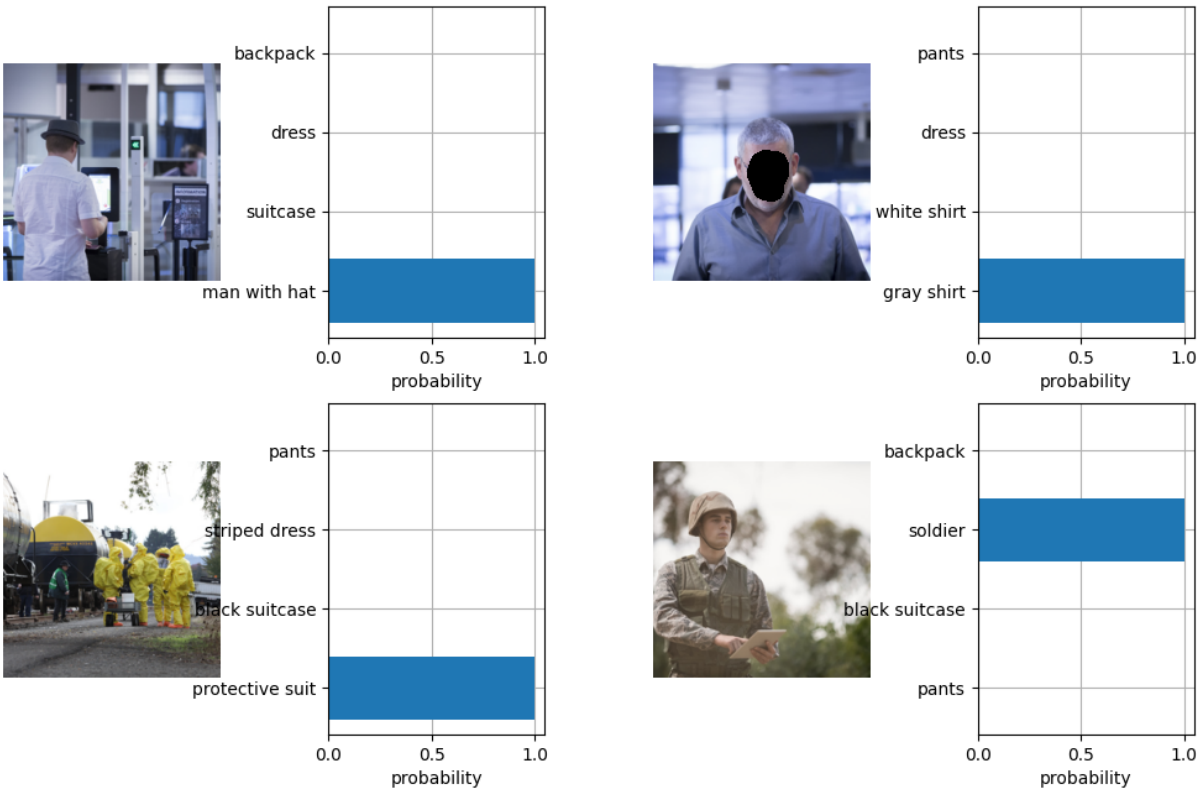

Figure 7. Sample results of CLIP model. Most likely attribute has highest probability.

ume 33, pages 8295–8302, 2019. 2

[11] R. Gong, X. Liu, S. Jiang, T. Li, P. Hu, J. Lin, F. Yu, and J. Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4852–4861, 2019. 2

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 4

[13] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[14] K. Han, J. Guo, C. Zhang, and M. Zhu. Attribute-aware attention model for fine-grained representation learning. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 2040–2048, 2018. 2

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 6

[16] C. Huang, Y. Li, C. C. Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. 2

[17] J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang. Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method. *arXiv preprint arXiv:2005.11909*, 2020. 3

[18] F. M. Khan and F. Bremond. Unsupervised data association for metric learning in the context of multi-shot person re-identification. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 256–262. IEEE, 2016. 1

[19] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 5

[20] B. Kulis et al. Metric learning: A survey. *Foundations and trends in machine learning*, 5(4):287–364, 2012. 2

[21] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 111–115. IEEE, 2015. 2

[22] D. Li, Z. Zhang, X. Chen, and K. Huang. A richly annotated pedestrian dataset for person retrieval in

real surveillance scenarios. *IEEE transactions on image processing*, 28(4):1575–1590, 2019. 2, 3, 5

[23] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016. 3, 4

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[25] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019. 1, 2

[26] H. Liu, J. Wu, J. Jiang, M. Qi, and B. Ren. Sequence-based person attribute recognition with joint ctc-attention model. *arXiv preprint arXiv:1811.08115*, 2018. 2

[27] P. Liu, X. Liu, J. Yan, and J. Shao. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102*, 2018. 2

[28] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 2

[29] X. Liu, J. Wang, S. Wen, E. Ding, and Y. Lin. Localizing by describing: Attribute-guided attention localization for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 2

[30] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017. 2, 3

[31] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 3

[32] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th international conference on pattern recognition*, volume 1, pages 582–585. IEEE, 1994. 1

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015. 5

[34] H. Qin, R. Gong, X. Liu, M. Shen, Z. Wei, F. Yu, and J. Song. Forward and backward information retention for accurate binary neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2250–2259, 2020. 2

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 5

[36] A. Santos, C. Ortiz de Solórzano, J. J. Vaquero, J. M. Pena, N. Malpica, and F. del Pozo. Evaluation of autofocus functions in molecular cytogenetic analysis. *Journal of microscopy*, 188(3):264–272, 1997. 3

[37] N. Sarafianos, X. Xu, and I. A. Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 680–697, 2018. 2

[38] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2, 5

[39] P. Sudowe, H. Spitzer, and B. Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 87–95, 2015. 2

[40] C. Sun, N. Jiang, L. Zhang, Y. Wang, W. Wu, and Z. Zhou. Unified framework for joint attribute classification and person re-identification. In *International Conference on Artificial Neural Networks*, pages 637–647. Springer, 2018. 2

[41] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015. 2

[42] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898, 2014. 5

[43] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 5, 6

[44] C. Tang, L. Sheng, Z. Zhang, and X. Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4997–5006, 2019. 2

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[46] G. Wang, J. Lai, P. Huang, and X. Xie. Spatial-temporal person re-identification. In *Proceedings of*

the AAAI conference on artificial intelligence, volume 33, pages 8933–8940, 2019. 2

[47] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393, 2014. 2

[48] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 2

[49] Y. Wu, Y. Wu, R. Gong, Y. Lv, K. Chen, D. Liang, X. Hu, X. Liu, and J. Yan. Rotation consistent margin loss for efficient low-bit face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6866–6876, 2020. 2

[50] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[51] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 5

[52] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39. IEEE, 2014. 1

[53] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013. 4

[54] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Jin. Grouping attribute recognition for pedestrian with joint recurrent learning. In *IJCAI*, pages 3177–3183, 2018. 2

[55] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015. 4

[56] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[57] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019. 4

[58] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017. 4