ASMR: Learning Attribute-Based Person Search with Adaptive Semantic Margin Regularizer — Supplementary Materials —

Boseung Jeong1*Jicheol Park2*Suha Kwak1,2Dept. of CSE, POSTECH1,Graduate School of AI, POSTECH2

http://cvlab.postech.ac.kr/research/ASMR/

1. Architecture Details

This section describes details of our model architecture that consists of two encoders, *image encoder* and *person category encoder*. Configurations of the two encoders are elaborated in Table 1, where d_{pc} denotes the dimension of a person category vector, a binary vector obtained by concatenating one-hot vectors of its all attributes. Specifically, d_{pc} is 105, 30, and 26 for the PETA [2], Market-1501 Attribute [3], and PA100K [4] datasets, respectively.

Image	encoder	Person category encoder			
Structure	Size	Structure	Size		
ResNet-50	2048				
+ GAP	2040				
FC.	2048×512	FC.	$d_{pc} \times 512$		
101	ReLU	101	ReLU		
FCa	512×128	FCa	512×128		
102	ReLU	102	ReLU		
FC ₃	128×128	FC_3	128×128		

Table 1. Details of the two encoders.

2. Pretraining of Image Encoder

Before training of our model, the image encoder is pretrained by multiple attribute classifiers to make the image representation more suitable to person search. As shown in Figure 1, we append a classifier head with four Fully Connected (FC) layers on top of Global Average Pooling (GAP) for each attribute group. Each classifier is learned to choose the correct attribute among those in each attribute group. It consequently improves the representation power of the image encoder backbone. Specifically, the image encoder is pretrained by the softmax cross-entropy loss per classifier:

$$\mathcal{L}_{cls} = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} \log(p(I_i, y_{ij})), \qquad (1)$$



Figure 1. Pretraining of the image encoder. Cls and CE loss denote the attribute classifiers and Softmax Cross Entropy loss, respectively. After pretraining, the multi-label classification part is removed and only the CNN backbone is used for the next stage.

where $p(I_i, y_{ij})$ means the predicted probability of the *i*-th training image I_i for its groundtruth attribute of the *j*-th attribute group y_{ij} while *m* and *n* denote the number of training images and that of attribute groups, respectively. The classification loss \mathcal{L}_{cls} is applied at the pretraining stage only, and the pretrained CNN is utilized as the backbone of the image encoder at the next stage.

3. Details about Attribute Groups

A person category is represented as a binary vector, which consists of exclusive attribute groups such as Gender, Age and Accessory. We define and utilize the attribute groups for both pretraining of the image encoder and calculating the weighted Hamming distance. The attribute groups of each dataset we adopt are enumerated in Table 2.

4. Effect of a Hyper-parameter

To investigate the effect of ASMR, an ablation study is conducted by varying the value of λ , the importance weight for ASMR in Eq. (1) of the main paper, on the three datasets. As shown in Table 3, ASMR improves performance on all the three datasets when $\lambda \ge 4$; these results suggest that ASMR is effective regardless of datasets if its importance weight is sufficiently large. We also stress that, as we tune λ for each dataset to obtain optimal scores, the state of the art (*i.e.*, SAL [1]) also tune multiple hyper-

^{*}Equal contribution

Dataset	Attribute group
	Age, Carrying, Upper body casual, Lower body casual, Accessory, Footwear,
PETA [2]	Kind of upper body, Sleeve, Kind of lower body, Texture of upper body, Texture of lower body,
	Gender, Hair length, Color of upper body, Color of lower body, Color of footwear, Color of hair
Montrat 1501 Attribute [2]	Age, Bag, Color of lower body clothing, Color of upper body clothing, Type of lower body clothing,
Market-1301 Attribute [5]	Length of lower body clothing, Sleeve length, Hair length, Hat, Gender
DA 100K [4]	Age, Gender, Viewpoint, HandBag, ShoulderBag, Backpack, HoldObjectsInFront, Hat, Glasses
FA100K [4]	Length of Sleeve, Patterns of upper body, Patterns of lower body, Coat, Kind of lower body, Boots

Table 2. Lists of attribute groups in the three benchmark datasets for attribute-based person search.

λ	0	1	2	3	4	5	6	7
PETA	48.5	49.5	50.5	50.5	56.5	53.0	51.5	52.0
Market	44.8	44.4	44.8	45.0	46.3	47.7	49.6	45.3
PA100K	28.9	28.2	27.9	28.9	30.7	31.9	30.2	29.0

Table 3. Performance in Rank@1 versus λ on the three datasets.

parameters (*e.g.*, learning rates for the attribute and image encoders) differently for different datasets to achieve their final scores. We further conduct an ablation study by varying σ and γ , the scale and margin for MA loss in Eq. (2) of the paper, on the PETA dataset. Fig. 2 demonstrates that our method consistently outperforms state of the art even with diverse hyper-parameter setting.

5. More Comparison to SAL

Our method outperforms SAL in terms of Rank-1 and mAP, yet worse in terms of Rank-5 and Rank-10. This implies that our retrieval results are more precise (Rank-1), and are less sensitive to the threshold of CMC (mAP). In addition, we reproduced SAL [1] through its official implementation¹ to compare our method with SAL in terms of training complexity and performance on the PA100K dataset and qualitative comparison. As shown in Table 4, the proposed method outperforms both of SAL and its reproduced version (SAL^{\dagger}) on all the three datasets in Rank1 and mAP. Moreover, we would stress that SAL was not stable in training and the records of SAL reported in the paper were not well reproduced; we suspect that the main reason for this failure would be the complicated training procedure of SAL based on adversarial learning. Lastly, Fig. 3 shows the further comparisons between ours and SAL[†] in terms of the training convergence (left) and the qualitative comparisons (right), on the Market-1501 dataset.

6. Failure Cases

Even though ASMR considers semantic dissimilarity between person categories, it sometimes fails then query attributes are not visually well-distinguishable. Fig. 4 show examples of such failures due to subtle appearance differences between "adult" and "teenager" or between "bag" and "handbag".



Figure 2. Rank-1 versus σ and γ on the PETA dataset.

7. Qualitative Results

More qualitative results of our method on the three public datasets are presented in Fig. 5, 6, and 7. Results on the PETA and the Market-1501 Attribute datasets overall demonstrate that our method is insensitive to pose variations. Also, our method learns body pose variations on the PA100K dataset with viewpoint labels. In detail, individual results show that our method is robust against changes in image resolution (Fig. 5(a,d-j), Fig. 6(b,c,f), Fig. 7(e-i)), illumination (Fig. 5(b,f-m), Fig. 6(b,c), Fig. 7(a-c)), and partial occlusions (Fig. 5(b,c,l), Fig. 7(l)). Even though some attributes are associated with tiny details of images such as hat (Fig. 5(n), Fig. 7(i,j)), bag (Fig. 5(b-e,m-o), Fig. 6(e-k), Fig. 7(e-k)), footwear (Fig. 5(d-o)), accessory (Fig. 5(e), Fig. 6(c,d)), and clothes patterns (Fig. 6(a-f, o)), our method well captures such fine details and retrieves images accurately. However, our method sometimes fails when query attributes are not visually well-distinguishable (Fig. 5(b,g,j,o), Fig. 6(f,h,o), Fig. 7(k,l)).

Finally, in Fig. 8, 9, and 10, we visualize the embedding manifold learned by our method through t-SNE on the test splits of the three datasets. The visualization results demonstrate that for most images their nearest neighbors are similar with them in terms of their appearance traits. This suggests that our method learns a semantic relation between

¹https://github.com/ycao5602/SAL



	teenager	-	pants	short	short sleeve
	short hair	male	up green	down blue	-
ASMR				SAL	

Figure 3. Further comparisons between ours and SAL. (*left*) Rank-1 versus training epoch on the Market-1501 dataset. (*right*) Top 3 retrieval results on the Market-1501 dataset.

adult	handbag	pants	shorts	short sleeve	adult	bag	dress	shorts	short sleeve
short hair	male	no	down black	-	long hair	female	up white	e down white	hat
o dult	bandbag	drocs	shorts	short	a duilt	bag	nante	long lower	
adult	Hanubag	uless	5110115	sleeve	aduit	Dag	pants k	ody clothing	
long hair	female	no	no	sleeve -	short hair	female	no	oody clothing down gray	-

Figure 4. Failure cases of our method on the Market-1501 Attribute dataset. Images are sorted from left to right according to their ranks. Green and red boxes indicate true and false matches, respectively. Queries are presented above their retrieved images; blanks indicate attributes that do not exist in the query. Colored red in query indicates attributes that are different between query person category and person categort of false matches.

Method	PET	ΓA	Market-150	PA100K		
	Rank1	mAP	Rank1	mAP	Rank1	mAP
SAL	47.0	41.2	49.0	29.8	-	-
SAL^{\dagger}	39.0	37.2	44.4	29.4	22.7	15.0
Ours	56.5	50.2	49.6	31.0	31.9	20.6

Table 4. Comparison of SAL, its reproduction by the official implementation ([†]), and our method on the three public datasets.

images and person categories successfully through the proposed loss.

References

 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. 1, [2] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proc. ACM Multimedia Conference (ACMMM)*, 2014. 1, 2

- [3] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 2019. 1, 2
- [4] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

Age <30	female	-	-	Age <30 m	nale backpac	k -	Age <30	male	messenger b	ag -
trousers	up red	down black	-	trousers up	black down bla	ck -	trousers	up black	down gray	-
1	1		I	and a					ġ	
		(a)			(b)				(c)	
Age <45	female	messenger	bag -	Age <30 ma	ale messenger	bag muffler	Age <45	male	-	-
trousers	up red	down bla	ck boots	trousers up b	lack down brov	vn leather shoes	trousers	up white	down gray	shoes
Î	*	Ŕ)	P	X	Ŷ	R	A		
		(d)			(e)				(f)	
Age <30	male	-	-	Age <45 m	iale -	-	Age <45	male	-	-
shorts	up black	down white	shoes	trousers up v	white down brov	wn leather shoes	jeans	up gray	down gray	shoes
				Â		Ì				
		(g)			(n)				(1)	
Age <45	female	-	-	Age <45 m	iale -	-	Age <45	female	-	-
jeans	up pink	down blue	sneaker	suits up	blue down blac	ck leather shoes	-	up black	down blue	shoes
		(j)	Â		(k)				(1)	
Age <45	male	messenger	bag -	Age >60 fe	male other	hat		female	backpac	.k -
jeans	up greer	n down blu	ue sneaker	trousers up	green down gre	en leather shoes	jeans	up black	down gre	ey sneaker
Ŕ		A							\$	
		(m)			(n)				(0)	

Figure 5. Top 5 retrieval results of our method on the PETA dataset. Images are sorted from left to right according to their ranks. Green and red boxes indicate true and false matches, respectively. Queries are presented above their retrieved images; blanks indicate attributes that do not exist in the query.

Age18-60	male	short sleev	2 -	Age18-60	male	short sleeve	-	Age18-60	male	long sleeve	-
-	side	upper plaid	trousers	-	back	upper plaid	trousers	glasses	side	upper plaid	trousers
		(a)			İ	(b)			R	(c)	
Age18-60 r	nale l	ong sleeve		Age18-6	0 mal	e short sleev	ve hand bag	Age18-60) mal	e short sleev	ve backpack
glasses :	side u	pper splice	trousers		sid	e upper log	o trousers	-	fror	nt upper log	o trousers
		(d)				(e)			R	(f)	
Age18-60 fe	male sh	ort sleeve sł	noulder bag	AgeOver6	50 ma	e short sleev	/e hand bag	Age18-60	male	long sleeve	backpack
- k	back	-	trousers		bac	k -	trousers	-	back	-	trousers
		(g)		Ş	Ŷ	(h)			P	(i)	
Age18-60 m	nale sho	ort sleeve sh	oulder bag	Age18-6	0 fema	ale long coat	hand bag	Age18-60	female	short sleeve	-
- fr	ont	-	trousers	-	bac	.k -	trousers	-	front	-	skirt&dress
		(j)			P	(k)				(1)	
AgeLess18	female	short sleev		AgeLess1	8 fema	e long sleeve	_	Age18-60	female	short sleeve	_
-	front	-	trousers	-	fron	t -	trousers	-	side	upper stride	skirt&dress
		5									

Figure 6. Top 5 retrieval results of our method on the PA100K dataset. Images are sorted from left to right according to their ranks. Green and red boxes indicate true and false matches, respectively. Queries are presented above their retrieved images; blanks indicate attributes that do not exist in the query.

(n)

(0)

(m)

Adult	-	pants	long lower body clothing	short sleeve	Teenager	-	pants	short	short sleeve
short hair	male	up blue	down black	-	short hair	male	up blue	down pink	-
								ý ž 🍘 ý	
			(a)					(b)	
Adult	-	pants	long lower body clothing	short sleeve	Adult	-	pants	long lower body clothing	short sleeve
short hair	male	up black	down brown	-	short hair	male	-	down black	-
									A A
Teenager ba	ckpack	pants	long lower body clothing	short sleeve	Adult	backpa	ck pant	s short	short sleeve
long hair f	emale	up white	e down black	-	short hair	male	up pur	ple down black	-
ST Y		Ŷ	¥£\$						
			(e)					(f)	
Teenager ba	ackpack	pants	long lower body clothing	short sleeve	Teenager	backpa	ck pant	s short	short sleeve
short hair	male	up yellov	w down green	-	short hair	female	e up yell	ow down gray	-
		Ŷ							
			(g)					(h)	
Teenager ba	ackpack	pants	long lower body clothing	short sleeve	Teenager b	backpacl	pants	short	short sleeve
long hair f	emale	up red	down black	hat	short hair	male	up whit	e down pink	hat
			(1)					()	
Adult ł	nandba	g dress	long lower body clothing	short sleeve	Adult	-	pants	long lower body clothin	g short sleeve
short hair	female	up red	down black	-	short hair	male	up red	down black	-
							K		

Figure 7. Top 10 retrieval results of our method on the Market-1501 Attribute dataset. Images are sorted from left to right according to their ranks. Green and red boxes indicate true and false matches, respectively. Queries are presented above their retrieved images; blanks indicate attributes that do not exist in the query.



Figure 8. 2D *t*-SNE visualization of image embeddings in the gallery of PETA.

Figure 9. 2D *t*-SNE visualization of image embeddings in the gallery of PA100K.

Figure 10. 2D t-SNE visualization of image embeddings in the gallery of Market-1501 Attribute.