Unsupervised Person Re-Identification with Transformer-based Network for Intelligent Surveillance Systems

Ge Cao
School of Electrical Engineering
University of Ulsan
Ulsan, Republic of Korea
caoge@islab.ulsan.ac.kr

Kang-Hyun Jo
School of Electrical Engineering
University of Ulsan
Ulsan, Republic of Korea
acejo@ulsan.ac.kr

***Abstract*—Person re-identification (re-ID) is an important topic in computer vision. In this paper, we study the unsupervised person re-ID which aims to identify target identity across multiple non-overlapping cameras for intelligent surveillance systems. The main challenge of unsupervised person re-ID lies in how to learn discriminative features without leveraging any annotated data. In this paper, we apply the Vision Transformer (ViT) to unsupervised person re-identification (re-ID) task. Combined with Multi-label Classification, the performance outperforms most CNN-based methods. We evaluate the proposed model on *Market-1501*, *DukeMTMC-reID* and *MSMT17* and achieves 56.6%, 49.4%, 14.5% in mAP, respectively, which outperforms the baseline by a clear margin and achieves the state-of-the-art unsupervised re-ID methods.**

***Index Terms*—Intelligent surveillance systems, Vision Transformer, Unsupervised learning**

I. Introduction

Person Re-Identification (re-ID) is a challenging task which aims to identify quantities of target images in a cross-camera system. For intelligent security and surveillance systems, person re-ID is the foundation of a wide range of applications, such as person tracking [1] and human activity analysis [2]. Existing person re-ID works mostly focus on supervised learning [3], [4] with the large-scale data-based annotated datasets. Due to the expensive cost for annotating a large number of person images across multiple cameras, recent researchers paid more attention to unsupervised person re-ID, which performs steadily where only unlabeled data is available.

CNN-based models [5], [6] have achieved great performance in unsupervised learning re-ID task. Nevertheless, if past development trend to go by, CNN-based models will sink into a hard period. After the work by Mnih *et al.* [7], attention mechanism related research became fashionable. As Transformer [8] became the most effective method in Natural Language Processing (NLP), many Transformer-based variants became popular and achieves the state-of-the-art in numerous computer-vision tasks, such as object recognition [9], object detection [10] and semantic segmentation [11]. The Vision Transformer (ViT) [9] is the first work to show even pure transformer architecture could achieve impressive performance in large-scale datasets for image classification.

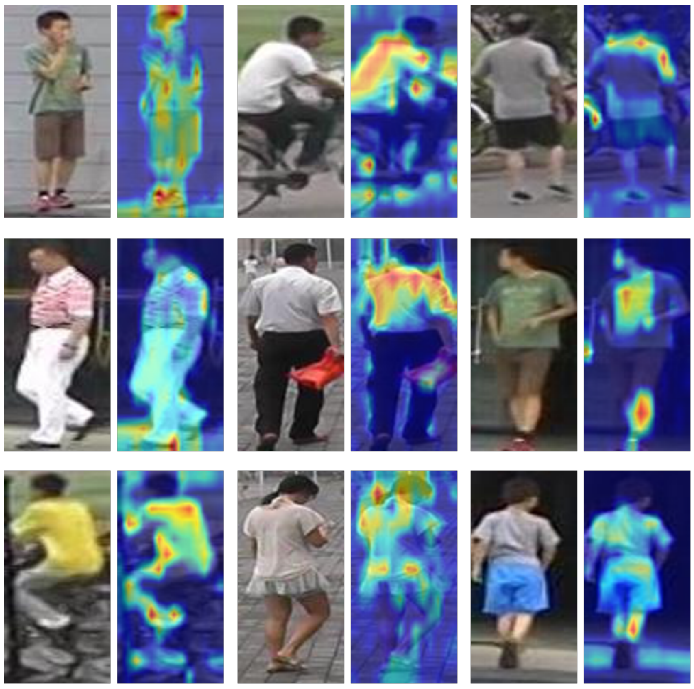

Fig. 1. Qualitative results of the proposed person re-ID model with attention map.

He *et al.* [4] applied a transformer-based model in supervised learning and outperformed existing CNN-based methods with a clear margin. To prove the effectiveness of pure transformer architecture in unsupervised learning, we follow the above works' experience and apply the transformer-based method into baseline MMCL [6]. The baseline we chose in this paper is proposed by Wang *et al.* [6], which gained state-of-the-art performance in the CNN-based method.

In this paper, we proposed an effective model which combines the MMCL training strategy and pure transformer architecture, the architecture is shown in Fig. 2. We conduct

an extensive comparison with the related methods which demonstrate the effectiveness of the proposed model in finding discriminative features. The visualization results is shown in Fig. 4. Our model achieves the state-of-the-art performance on the benchmark datasets Market-1501 [28], DukeMTMC-reID [29] and MSMT17 [30].

II. Related Work

This section briefly reviews related works on unsupervised person re-ID, multi-label classification, and transformer in vision.

A. Unsupervised Re-Identification

Unsupervised person re-ID works can be classified into three categories. The first category can be regarded as a collection of traditional unsupervised person re-ID methods hand-craft features design [12], localized salience statistic exploit [13] and dictionary learning based methods [14]. However, it is difficult to design robust and discriminative features in cross-camera conditions. And different illumination and viewpoint would hinder these kinds of methods to pursue better performance.

The second category used CNN models and adopted pseudo labels to cluster the training images. Typical representative is proposed by Yang *et al*. [15] which mined the potential similarity of unlabeled samples and achieved great success. Apart from that, Lin *et al*. [16] innovatively proposed a bottom-up clustering framework that trains network with pseudo labels generated by unsupervised learning iteratively. The third category applies transfer learning to deal with the problem of unsupervised person re-ID. ECN [17] proposed three types of underlying invariance to reduce feature distribution gap between the source and target domains. MAR [18] adopted a different strategy and used the source dataset as a reference to learn soft labels as ground truth to supervise the re-ID training. The baseline MMCL [6] we adopted in this paper is different from the three kinds of categories. It does not need any label dataset as an auxiliary reference and achieves better performance than most transfer learning methods.

B. Multi-label Classification

Multi-label classification is designed for recognition tasks to make multi-label as ground truth to supervise the training. Zhang *et al*. [19] summarized and reviewed the development of multi-label learning. Lin *et al*. [20] made the source and target datasets share the same set of mid-level semantic attributes. Similarly, Wang *et al*. [21] used multi-label classification to learn attribute features. Durand *et al*. [22] used GNN to predict missing labels in the training and generated partial labels to deal with the problem of multi-label learning.

C. Transformer in Vision

Attention aims to make networks tend to focus on important features and suppress irrelevant features. Many works learn attention using CNN with special residual operations. SENet [23] proposed squeeze-and-excitation connection which trained the spatial-wise attention. CBAM [24] further proposed channel-wise attention. In person re-ID task, RGA-SC [25] proved the effectiveness of the above attention method when applied in supervised training.

As a successful application of self-attention, Transformers were proposed by Vaswani *et al*. [8] for natural language processing. Then many studies have shown its effectiveness for computer tasks. Han *et al*. [26] has reviewed the application of the Transformer in computer vision field. Then ViT [9] recently applied a pure transformer architecture directly to sequences of image patches in image recognition. To overcome the shortcoming that ViT requires large-scale datasets to pretrain, Touvron *et al*. [27] proposed DeiT which speeds up ViT training by a teacher-student strategy. For supervised re-ID task, He *et al*. [4] proposed a pure-transformer architecture TransReID and got the superior promising performance. In this paper, we continue the style of ViT and apply a pure-transformer network for unsupervised person re-ID task.

III. Methodology

Our person re-ID In this section, we summarize the problem formulation and overview for unsupervised person reID task in Sec III.A. The training strategy MMCL [6] (baseline) is shown in Sec III.C. Then in Sec III.C, the transformer architecture we applied as backbone is introduced.

A. Problem formulation and Overview

Given an unlabeled dataset $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, where n means the number of identities, our goal is to train a person re-ID model on $\mathcal{X}$. The unlabeled dataset $\mathcal{X}$ consists of three subset: training set $\mathcal{T}$, query set $\mathcal{Q}$ and gallery set $\mathcal{G}$. When giving any query image q from query set $\mathcal{Q}$, we expect the re-ID model produce a feature vector f_i to retrieve image g in gallery set $\mathcal{G}$ containing the same identity. As shown in Fig. 2, when giving any input image x_i, we get a d-dimentional L2-normalized feature vector f_i extracted by the backbone, where $d = 2048$ and $\|f_i\| = 1$ in this paper.

To ensure the training on $\mathcal{X}$ possible, MMCL generates the single-class label y_i for each input image x_i to convert the question to a supervised case, where only the value at index i is set to 1 and the others are -1, *i.e.*,

$$y_i[j] = \begin{cases} 1 & j = i \\ -1 & j \neq i \end{cases} \tag{1}$$

Since for every person x_i in $\mathcal{X}$, there are multiple images containing the same identity index, so single-class label obviously cannot satisfy the requirement of unsupervised learning task. In next subsection, we would introduce how to produce multi-class label and the training strategy MMCL proposed by [6] in detail.

B. Memory-based Multi-label Classification

Label prediction requires multi-class label for each image. For any input image x_i, we can get 2048-d feature vector

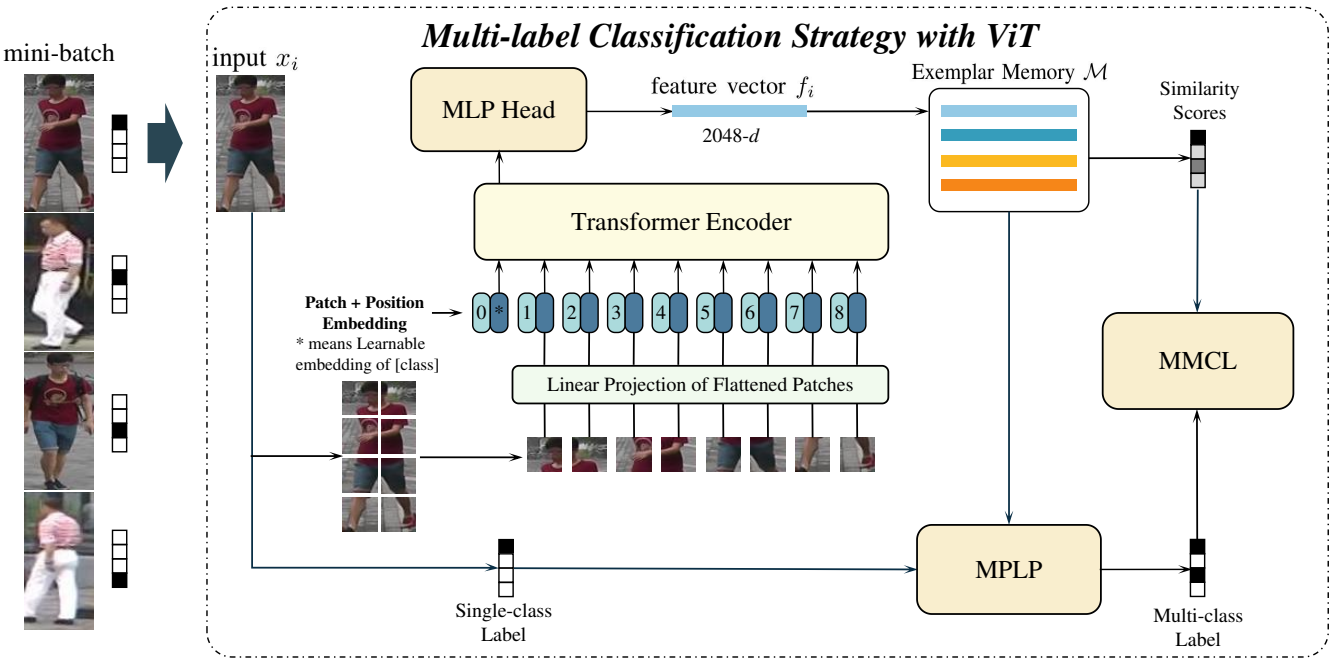

Fig. 2. The overview of the proposed model for unsupervised person re-ID. For each input x_i in target set $\mathcal{T}$, the model generates multi-class label by MPLP and runs MMCL for multi-class classification. Before entering the backbone, we apply LBP to augment the input image x_i and then split it into fixed-size patches. After linearly embedding and combining with position embeddings, the patches are fed as sequence of vectors into standard transformer encoder.

extracted by the backbone. So for images x_i and x_j, f_i and f_j can be used to compute the similarity score as,

$$s_{i,j} = f_i^{\mathsf{T}} \times f_j \tag{2}$$

Then MMCL [6] proposed a $n \times d$ Exemplar Memory $\mathcal{M}$ to store the feature vectors, where $\mathcal{M}_i = f_i$, and Memory-based Positive Label Prediction (MPLP) for generating multi-class labels $\bar{y}_i$, which requires the single-class label y_i and Exemplar Memory $\mathcal{M}$, *i.e.*,

$$\bar{y}_i = MPLP(y_i, \mathcal{M}) \tag{3}$$

The target of MPLP is to find other images which contain the same identity with x_i in $\mathcal{X}$. For $\mathcal{X}$, we have n feature vector $\{f_i | i = 0, 1, ..., n\}$. So we can get the similarity score $s_{i,j}$ between x_i and x_j by Eq. (2). For x_i, MPLP first computes a rank list R_i according to the similarity between x_i and other images, *i.e.*,

$$R_i = argsort_j(s_{i,j}), j = 1, 2, ..., n \tag{4}$$

R_i shows the similarity between x_i and others, but for the training, we need to judge the positive pairs and negative pairs. MMCL deals with it by apply a threshold t to remove the label with similarity smaller than t, then we get a new rank list P_i.

$$P_i = R_i[1 : k_i] \tag{5}$$

where k_i denotes there are k_i label candidates can be regarded as similar pair with x_i. Then inversely for x_j in P_i, if label i also one of the top-k_i labels of x_j, x_j is considered as a positive pair with x_i. So the positive label set can be denoted as,

$$P_i^* = p_i[1 : l] \tag{6}$$

where l satisfies $i \in R_{P_i[l]}[1 : k_i]$ & $i \notin R_{P_i[l+1]}[1 : k_i]$, so multi-class label $\bar{y}_i$ is,

$$\bar{y}_i[j] = \begin{cases} 1 & j \in P_i^* \\ -1 & j \notin P_i^* \end{cases} \tag{7}$$

In this paper, we follow the Memory-based Multi-label Classification Loss (MMCL) as,

$$l^*(j|x_i) = \left\| \mathcal{M}[j]^{\mathsf{T}} \times f_i - \bar{y}_i[j] \right\|^2 \tag{8}$$

C. Backbone: Vision Transformer

The architecture of the backbone model ViT is depicted in Fig. 2. The standard Transformer receives the input tensor as $x_i \in \mathbb{R}^{H \times W \times C}$, where (H, W) is the resolution of the input image, C is the number of channels. Follow the mode of ViT [9], we reshape the image x_i into a sequence of flattened 2D patches $p_i \in \mathbb{R}^{N \times (P^2 C)}$, where (P, P) is the resolution of each patch, so $N = HW/P^2$ denotes the number of patches. The Transformer Encoder requires 1D vertor through all its layer, so we convert the input tensor into D-dimensional vector with a trainable linear projection. As shown in Fig. 3(b), we apply a convolutional layer to convert input tensor into $p_i \in \mathbb{R}^{(H/P \times W/P) \times D}$.

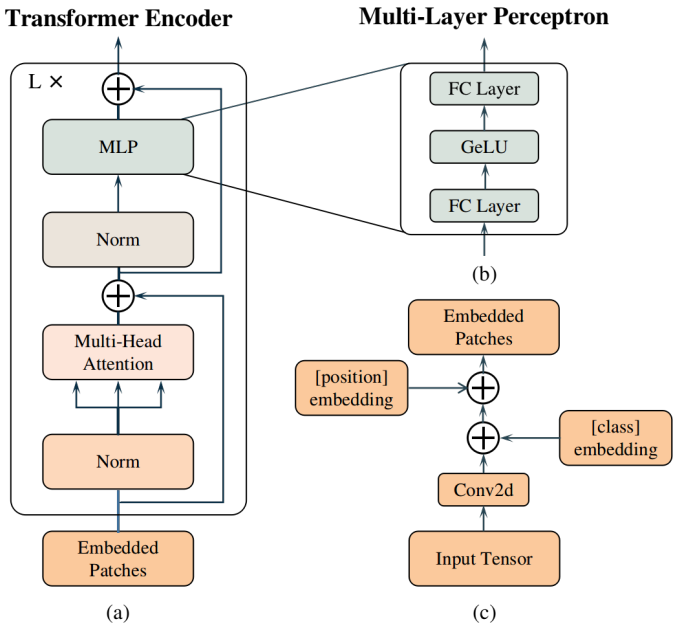

Fig. 3. Detail architecture of Transformer Encoder. (a) Architecture of Transformer Encoder. (b) Architecture of Multi-Layer Perceptron (MLP). (c)The process of producing embedded patches from input tensor.

Different with original Transformer [8], a learnable embedding [class] embedding $c_i \in \mathbb{R}^{1\times D}$ is concatenated to p_i to get $p'_i \in \mathbb{R}^{(H/P\times W/P+1)\times D}$. Then learnable [position] embeddings are added to the patches embedding p_i to generate the Embedded Patches $z_i \in \mathbb{R}^{(HW/P^2+1)\times D}$ at the bottom of Fig. 3(a).

Embedded patches z_i is the input of Transformer Encoder depicted in Fig. 3(a). The encoder is composed of Multi-Head self-attention module (MSA) and Multi-Layer Perceptron (MLP). Layernorm operation (LN) is applied before MSA and MLP, and MLP contains a GELU non-linearity and dropout. The Transformer Encoder is circularly applied L times in the whole backbone, i.e.,

$$z'_{i,l} = MSA(LN(z_{i,l-1})) + z_{i,l-1}, \qquad l = 1, 2, ..., L \quad (9)$$

$$z_{i,l} = MLP(LN(z'_{i,l})) + z'_{i,l}, \qquad l = 1, 2, ..., L \quad (10)$$

where z_i mentioned above is same to $z_{i,0}$. Then for the final classifier MLP Head, the input can be computed as,

$$y = LN(z^0_{i,L}) \quad (11)$$

To obtain the 2048-d feature vector, in this paper we change the $output_feature$ of MLP Head to 2048.

IV. Experiments

A. Dataset and Evaluation Metrics

Market-1501 [28], *DukeMTMC-reID* [29] and *MSMT17* [30] datasets are applied in this paper. *Market-1501* contains 32,668 labeled person images of 1,501 identities collected from 6 non-overlapping camera views. *DukeMTMC-reID* contains 36,411 annotated images of 1,404 identities with 8 cameras. *MSMT17* has 126,411 labeled person images of 4,101 with 15 cameras, which is the biggest released person re-ID dataset. These datasets are constructed with large amount of annotated images which collected from different view camera, illumination, indoor or outdoor scene and other variations. More details can be found in Table. 1. In this paper, we follow the standard settings with [6], [17]. We don't use any other labeled datset when training and testing and the performance is evaluated by Mean average precision (mAP) and the Cumulative Matching Characteristic (CMC) Rank-1/5/10 matching accuracy.

B. Implementation Details

We set the batch size to 32 in training and testing and all experiments are implemented on PyTorch. We use ViT [9] as backbone with the pre-trained model on ImageNet [34]. We change the final MLP Head's output feature from 768 to 2048 then it produces the 2048-d feature vector. The exemplar memory is initialized to all zeros and we start the MPLP for label prediction after 5 epochs. For data augmentation, we follow the methods of [6] and leverage CamStyle [35] as main data augmentation method. Other augmentations like random crop, random rotation, color jitter, and random erasing are also introduced to improve the robustness of the proposed model.

All the size of input images is resized to 224×224. The optimization method is SGD and the learning rate if 0.01 and weight decay is 0.0001. We train the model for 60 epochs, and the learning rate will divide by 10 every 40 epochs. The threshold t is set to 0.6 follow [6].

C. Comparison of the State-of-the-Art

We compare the proposed method against state-of-the-art unsupervised learning works on *Market-1501* [28], *DukeMTMC-reID* [29] and *MSMT17* [30]. The comparison results are shown in Table. 2 and Table. 3.

Table. 2 shows the comparisons on *Market-1501* and *DukeMTMC-reID*. We only compare with the methods which don't need any other labeled dataset: LOMO [12], BOW [28], UDML [31], DECAMEL [32], BUC [16], DBC [33] and MMCL [6].

In the comparison methods, LOMO and BOW used traditional unsupervised learning methods which utilizes hand-crafted features and got lower results compared with others. UDML proposed a multi-task dictionary learning method to learn dataset-shared but target-data-biased representation. DECAMEL, BUC and DBC use the clustering method to train their networks. And the MMCL is the baseline of this paper. It is obvious that our proposed model outperforms other works with a large margin. For example, we obtain 85.9% and 56.6% in Rank-1 and mAP on *Market-1501*, which surpass the baseline MMCL by 7.0% and 24.4%. On *DukeMTMC-reID*, we achieve 71.6% and 49.4% in Rank-1 and mAP, and outperform MMCL by 9.8% and 22.9%.

For comparison results of *MSMT17*, we compare with PTGAN [30], ECN [17], SSG [15] and the baseline MMCL [6]. Here we compare our model with some transfer learning

TABLE I
DETAILED INFORMATION OF DATASET MARKET-1501, DUKEMTMC-REID AND MSMT17.

Dataset	#ID	#ID detail			#image	#image detail			#cam
		Train	Query	Gallery		Train	Query	Gallery	
Market-1501 [28]	1,501	751	750	751	32,668	12,936	3,368	1,6364	6
DukeMTMC-reID [29]	1,404	702	702	1,110	36,411	16,522	2,228	17,661	8
MSMT17 [30]	4,101	1,041	3,060	3060	126,411	32,621	11,659	82,161	15

TABLE II
UNSUPERVISED PERSON RE-ID PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON MARKET-1501 AND DUKEMTMC-REID.

Method	reference	Market-1501					DukeMTMC-reID				
		Source	Rank-1	Rank-5	Rank-10	mAP	Source	Rank-1	Rank-5	Rank-10	mAP
LOMO [12]	CVPR15	None	27.2	41.6	49.1	8	None	12.3	21.3	26.6	4.8
BOW [28]	ICCV15	None	35.8	52.4	60.3	14.8	None	17.1	28.8	34.9	8.3
UDML [31]	CVPR16	None	34.5	52.6	59.6	12.4	None	18.5	31.4	37.6	7.2
DECAMEL [32]	TPAMI18	None	60.2	76	81.1	32.4	-	-	-	-	-
BUC [16]	AAAI19	None	66.2	79.6	84.5	38.3	None	47.4	62.6	68.4	27.5
DBC [33]	BMVC19	None	69.2	83	87.8	41.3	None	51.5	64.6	70.1	30
MMCL (Baseline) [6]	CVPR20	None	80.3	89.4	92.3	45.5	None	65.2	75.9	80	40.2
Ours	This paper	None	**85.9**	**92.3**	**94.3**	**56.6**	None	**71.6**	**80.8**	**83.9**	**49.4**

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS ON MSMT17.

Method	reference	Setting	MSMT17				
			Source	Rank-1	Rank-5	Rank-10	mAP
PTGAN [30]	CVPR18	UDA	Market	10.2	-	24.4	2.9
ECN [17]	CVPR19	UDA	Market	25.3	36.3	42.1	8.5
SSG [15]	ICCV19	UDA	Market	**31.6**	-	**49.6**	**13.2**
PTGAN [30]	CVPR18	UDA	Duke	11.8	-	27.4	3.3
ECN [17]	CVPR19	UDA	Duke	30.2	41.5	46.8	10.2
SSG [15]	BMVC19	UDA	Duke	**32.2**	-	**51.2**	**13.3**
MMCL(Baseline) [6]	CVPR20	Unsupervise	None	35.4	44.8	49.8	11.2
Ours	This paper	Unsupervise	None	**38.5**	**47.9**	**52.8**	**14.5**

methods because the pure unsupervised learning method is too little. Our model achieves 38.5% and 14.5% in Rank-1 and mAP respectively. It outperforms the baseline MMCL by 29.5% in mAP and other transfer learning methods even they use other auxiliary labeled datasets.

D. Visualization

Due to the different architecture with CNN-based methods, we cannot use the same visualization tool like Grad-CAM [36] which applied in many works to check the visualized results. Similar to ViT, we also use the tool Gradient Attention Rollout [37]. For the options of $head - fusion$, we choose maximum the attention weights, which is different with ViT, to get the results. The two kinds of tools can identify the regions that the network considers important. In Fig. 4, there are four groups' images of results. In each group, the first column is the input image, and the second is the attention map generated by the baseline MMCL [6], the third one is produced by the proposed model. In Fig. 4(a) and (b), It is obvious that when the proposed model processes the images with a bicycle inside, the proposed model still could focus on the person, which is the correct and discriminative feature of the identity. Compared with our model, the baseline would focus more on the background and unrelated objects. In Fig. 4(c), we can see that the attention also focuses on the red bag which is the most recognizable feature, the baseline MMCL

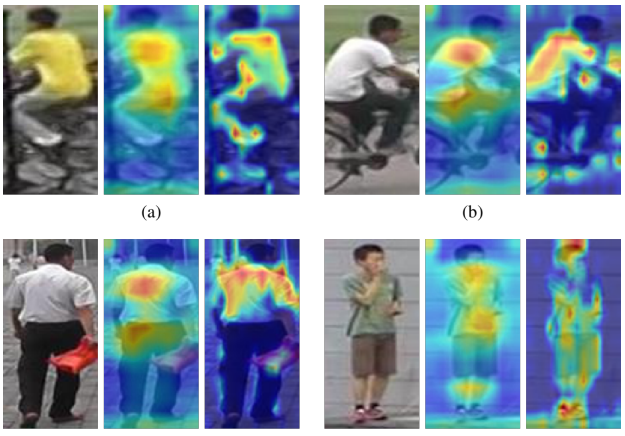

Fig. 4. Representative examples of attention from the output token to the input space. The first image is the input image taken from *Market-1501* dataset. The second column of each group is the attention map generated by the baseline. The third is produced by the proposed model.

just focuses on the whole person but no key-points on the most recognizable feature. In Fig. 4(d), the attention generated by the baseline MMCL diffuses to a large part of the background while our model can firmly focus on the person, even the edges of persons and background if clearly visible.

V. Conclusions

This paper proposes a multi-label classification method with pure Transformer architecture to address unsupervised person re-ID task. This work does not require any auxiliary annotated data and the performance outperforms most works applying transfer learning methods. The results on three public released datasets prove the robustness and effectiveness of the proposed model. The great improvement proves the infinite potential of the transformer in unsupervised person re-ID task. Furthermore, we can improve it from data augmentation, architecture variant and other aspects.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the government(MSIT).(No.2020R1A2C200897212)

References

[1] J. Nino, A. Frias Velazquez, N. Bo, M. Slembrouck, J. Guan, G. Debard, B. Vanrumste, T. Tuytelaars, and W. Philips, "Scalable semi-automatic annotation for multi-camera person tracking," *IEEE Transactions on Image Processing*, vol. 25, pp. 1–1, 05 2016.

[2] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-identification: Past, Present and Future," *arXiv e-prints*, p. arXiv:1610.02984, Oct. 2016.

[3] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-Scale Feature Learning for Person Re-Identification," *arXiv e-prints*, p. arXiv:1905.00953, May 2019.

[4] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based Object Re-Identification," *arXiv e-prints*, p. arXiv:2102.04378, Feb. 2021.

[5] P. Khorramshahi, N. Peri, J.-c. Chen, and R. Chellappa, "The Devil is in the Details: Self-Supervised Attention for Vehicle Re-Identification," *arXiv e-prints*, p. arXiv:2004.06271, Apr. 2020.

[6] D. Wang and S. Zhang, "Unsupervised Person Re-identification via Multi-label Classification," *arXiv e-prints*, p. arXiv:2004.09228, Apr. 2020.

[7] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent Models of Visual Attention," *arXiv e-prints*, p. arXiv:1406.6247, Jun. 2014.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv e-prints*, p. arXiv:1706.03762, Jun. 2017.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv e-prints*, p. arXiv:2010.11929, Oct. 2020.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *arXiv e-prints*, p. arXiv:2005.12872, May 2020.

[11] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers," *arXiv e-prints*, p. arXiv:2012.15840, Dec. 2020.

[12] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person Re-identification by Local Maximal Occurrence Representation and Metric Learning," *arXiv e-prints*, p. arXiv:1406.4216, Jun. 2014.

[13] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," *BMVC 2014 - Proceedings of the British Machine Vision Conference 2014*, 01 2014.

[14] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised 11 graph learning," vol. 9905, 10 2016, pp. 178–195.

[15] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. Huang, "Self-similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-identification," *arXiv e-prints*, p. arXiv:1811.10144, Nov. 2018.

[16] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8738–8745, 07 2019.

[17] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-identification," *arXiv e-prints*, p. arXiv:1904.01990, Apr. 2019.

[18] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised Person Re-identification by Soft Multilabel Learning," *arXiv e-prints*, p. arXiv:1903.06325, Mar. 2019.

[19] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, pp. 1819–1837, 08 2014.

[20] S. Lin, H. Li, C.-T. Li, and A. Chichung Kot, "Multi-task Mid-level Feature Alignment Network for Unsupervised Cross-Dataset Person Re-Identification," *arXiv e-prints*, p. arXiv:1807.01440, Jul. 2018.

[21] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Identification," *arXiv e-prints*, p. arXiv:1803.09786, Mar. 2018.

[22] T. Durand, N. Mehrasa, and G. Mori, "Learning a Deep ConvNet for Multi-label Classification with Partial Labels," *arXiv e-prints*, p. arXiv:1902.09720, Feb. 2019.

[23] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *arXiv e-prints*, p. arXiv:1709.01507, Sep. 2017.

[24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," *arXiv e-prints*, p. arXiv:1807.06521, Jul. 2018.

[25] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-Aware Global Attention for Person Re-identification," *arXiv e-prints*, p. arXiv:1904.02998, Apr. 2019.

[26] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A Survey on Visual Transformer," *arXiv e-prints*, p. arXiv:2012.12556, Dec. 2020.

[27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv e-prints*, p. arXiv:2012.12877, Dec. 2020.

[28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," 12 2015, pp. 1116–1124.

[29] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking," *arXiv e-prints*, p. arXiv:1609.01775, Sep. 2016.

[30] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-Identification," *arXiv e-prints*, p. arXiv:1711.08565, Nov. 2017.

[31] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1306–1315.

[32] H. Yu, A. Wu, and W. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 956–973, 2020.

[33] G. Ding, S. H. Khan, and Z. Tang, "Dispersion based clustering for unsupervised person re-identification," in *BMVC*, 2019.

[34] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[35] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera Style Adaptation for Person Re-identification," *arXiv e-prints*, p. arXiv:1711.10295, Nov. 2017.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *arXiv e-prints*, p. arXiv:1610.02391, Oct. 2016.

[37] S. Abnar and W. Zuidema, "Quantifying Attention Flow in Transformers," *arXiv e-prints*, p. arXiv:2005.00928, May 2020.

[38] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-Aware Global Attention for Person Re-identification," *arXiv e-prints*, p. arXiv:1904.02998, Apr. 2019.