2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) | 978-1-6654-6382-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/AVSS56176.2022.9959509

Attribute De-biased Vision Transformer (AD-ViT) for Long-Term Person Re-identification

Kyung Won Lee, Bhavin Jawade, Deen Mohan, Srirangaraj Setlur, and Venu Govindaraju

Department of Computer Science and Engineering, University at Buffalo

{klee43, bhavinja, dmohan, setlur, govind}@buffalo.edu

Abstract

Person re-identification (re-ID) aims to retrieve images of the same identity from a gallery of person images across cameras and viewpoints. However, most works in person re-ID assume a short-term setting characterized by invariance in appearance. In contrast, a high visual variance can be frequently seen in a long-term setting due to changes in apparel and accessories, which makes the task more challenging. Therefore, learning identity-specific features agnostic of temporally variant features is crucial for robust long-term person Re-ID. To this end, we propose an Attribute De-biased Vision Transformer (AD-ViT) to provide direct supervision to learn identity-specific features. Specifically, we produce attribute labels for person instances and utilize them to guide our model to focus on identity features through gradient reversal. Our experiments on two long-term re-ID datasets - LTCC and NKUP show that the proposed work consistently outperforms current state-of-the-art methods.

1. Introduction

Person re-identification (re-ID) research focuses on identifying individuals across different spatial and temporal points and across cameras. It has increasingly gained attention due to its applications in video surveillance [34], cross-camera person tracking [37], and multi-camera activity analysis [2]. Person re-ID is typically categorized into short-term and long-term scenarios. The former applies to images captured within a few minutes to hours with limited variations in appearance. In contrast, the latter scenario considers images captured over many days or months with more diverse appearance changes.

978-1-6654-6382-9/22/$31.00 ©2022 IEEE

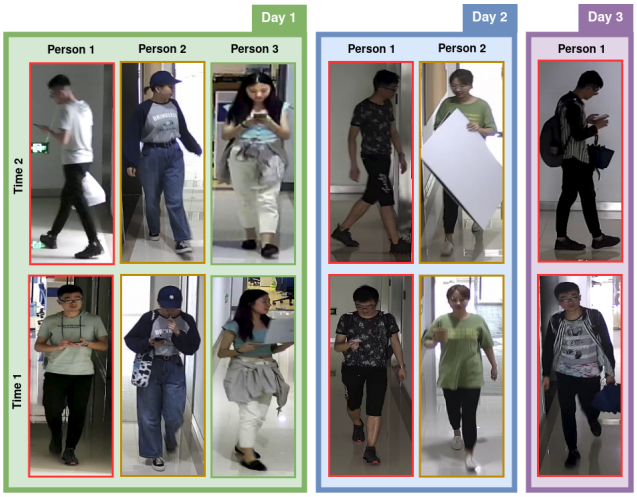

Figure 1: Example images of three subjects in the LTCC dataset. Person 1 and person 2 appear on more than one day with different outfits and accessories. Person 3 appears only for a day.

The research community has made significant advancements in the short-term scenario for more than a decade. Existing methods in short-term person re-ID have evolved from extracting hand-crafted features [8, 11] and distance metric learning [16, 20] to deep-learning-based methods [14, 29, 7, 25, 9] to solve challenges emerging from changes in viewpoints, occlusion, pose variations, misalignment, and varying illumination. This has resulted in robust re-ID performance on short-term benchmark datasets where the individual is wearing the same clothes across the data capture period.

In contrast, progress in the long-term scenario has been challenging, and there is significant room for improvement. For real-world applications, the long-term re-ID problem is of greater relevance. For instance, in many law enforcement-related applications, surveillance video

footage can span an extended period where the same individuals are likely to be seen with different clothes and apparel. Additionally, as shown in Figure 1, even in data spanning a short period of a few minutes to hours, individuals could put on or remove outerwear such as jackets, caps, sunglasses, etc., and carry or drop accessories such as cell phones, bags, laptops, etc. In other words, short-term person re-ID can also present some aspects associated with long-term re-ID.

To address the high variance in appearance that one observes in the long-term scenario, we propose an Attribute De-biased Vision Transformer (AD-ViT). Specifically, we apply a gradient reversal-based domain-adaptation mechanism over attribute cues to force the network to learn identity-specific features independent of apparel and accessory information.

In summary, our main contributions are as follows.

- We introduce a transformer-based framework called Attribute De-biased Vision Transformer (AD-ViT), which learns identity-specific features by utilizing person attribute information along with side information.
- We incorporate a gradient reversal mechanism based on adversarial learning to capture attribute-agnostic representations to address the long-term person re-ID problem.
- We perform experiments on two long-term cloth-changing re-ID datasets: LTCC [26] and NKUP [33]. Results demonstrate that our model consistently outperforms the baseline and the state-of-the-art in the long-term matching scenario.

2. Related Works

Short-term Person Re-ID In early research in the context of the short-term person re-ID setting, most methods focused on extracting person features through novel architectures [8, 11] and learning robust distance metrics [16, 20]. With the emergence of deep learning, various CNN-based methods have been investigated focused on challenges such as variations in camera viewpoints [14, 29], pose variation [7, 25], and occlusion [9]. Additionally, based on the intuition that different parts/regions of the image may contain useful cues for short-term re-ID, fine-grained features have also been learned [30, 32]. Since people tend to wear the same clothes over a brief period in the short-term setting, these approaches mostly employ global and local appearance features as the significant discriminating factor. However, appearance-based models fail when applied to the long-term re-ID setting because of considerable variations in clothing and accessories.

Long-Term Person Re-ID Lately, substantial attention has been given to the problem of long-term person re-ID, which primarily aims to learn clothing-agnostic cues from biometric modalities such as body shape [26, 12], motion [13, 1], and faces [31]. Qian et al. [26] employed identity-sensitive and clothing-insensitive representations using body keypoints. Hong et al. [12] proposed FSAM, which transfers fine-grained body shape knowledge to complement the clothing-agnostic knowledge while learning more discriminative human masks. Jin et al. [13] presented GI-ReID framework, which handles gait information as a regulator to facilitate the clothing-agnostic representation. Bansal et al. [1] introduced a vision transformer-based framework with gait-motion features as inputs to the multi-headed attention module. Wan et al. [31] employed a holistic and face feature extractor. However, some clothing-related representations have also been employed to address long-term person-ID tasks [18, 17]. Lee et al. proposed two sets of methods to identify people: (1) Clothing Model (CM) to identify clothing items with Wardrobe Model (WM) [18] to utilize these clothing attributes and defining a wardrobe set for an individual that it belongs to; and (2) Color Label Clothing Model (CL-CM) and Bayesian Personalized-Wardrobe Model (BP-WM) [17] to focus on incorporating an individual's preferential choices in attire. However, none of the previous works have explored de-biasing person attributes to supervise the network to extract features agnostic of clothing and accessories for long-term re-ID.

Domain Adaptation Domain adaptation methods try to supervise a feature extractor to learn representations that are agnostic of the source distribution domain and could generalize to a target distribution. Early works in this domain explored divergence-based domain adaptation, where the goal is to minimize some divergence criterion between the source and target distribution. Frequently utilized divergence measures include Correlation Alignment (CORAL) [28], Contrastive Domain Discrepancy (CDD) [15], and Maximum Mean Discrepancy (MMD) [27]. Some methods, such as DeepJDOT [3], have investigated optimal transport to minimize discrepancies in feature representations. Others [22, 36, 6] have investigated adversarial training methods where a generator and discriminator are utilized to learn domain-agnostic features. Ganin et al. [5] demonstrated that reversing the gradient of an auxiliary domain prediction objective with respect to a feature extractor's parameters aids in learning deep features that are domain-invariant. However, to the best of our knowledge, we are the first to apply domain adaptation methods to supervise the network to focus on identity-specific features, and to address long-term person re-ID.

3. Proposed Model

As illustrated in Figure 2, our model consists of two parts: (i) vision transformer based feature extractor and (ii)

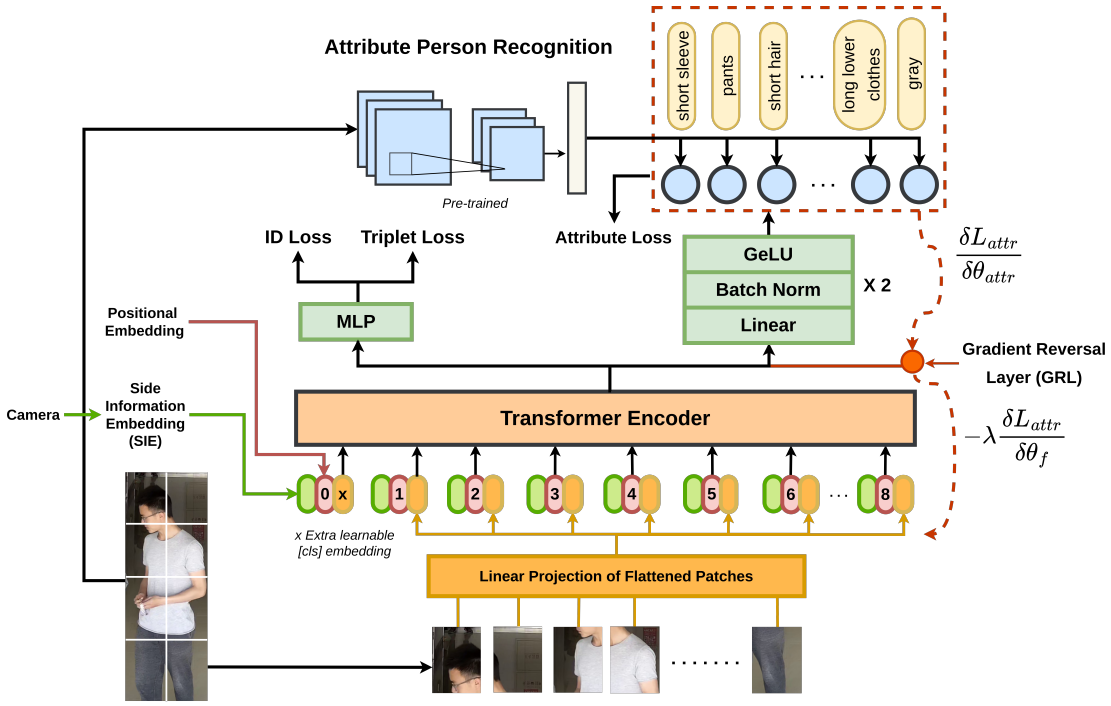

Figure 2: Framework of our proposed AD-ViT. Input images are split into eight image patches and then fed to the Transformer Encoder along with side information and positional embeddings after flattening. The final projection layer is included to capture global feature for subjects. Additionally, we use auxiliary projection layers to extract attribute features. We adopt Gradient Reversal Layer (GRL) for de-biasing attribute information by utilizing the predicted attribute labels.

an attribute de-biasing module. In Section 3.1, we describe the transformer-based baseline that we adopt for extracting subject-specific features. Section 3.2 explains the gradient reversal based attribute de-biasing architecture.

3.1. Baseline: Transformer-based Model

Let $I \in R^{H \times W \times C}$ be an image of a person where H, W, and C denote its height, width, and the number of channels. We split each image I into N patches, each of dimension M. Following [4], we flatten each patch and apply linear projection to extract a visual patch embedding. Let $V = \{v_1, v_2, ..., v_N\}$ be the visual patch embeddings corresponding to the original image.

Following [10], we add camera Side Information Embedding (SIE) to the patch embeddings. SIE embeddings are static embeddings of dimension $N_c \times D$, where N_c is the number of the camera and D is the embedding dimension.

$$Z = V + \lambda_s \cdot S_j \tag{1}$$

Here, λ_s denotes the SIE weighing factor, and S_j is the camera information embedding for camera ID 'j'. Finally, following [4], we add positional embeddings *pos* to all the patch embeddings along with an extra learnable [cls] embedding token. The final input sequence Z for the transformer can be represented as:

$$Z = \{x_{cls}; v_1 + \lambda_s \cdot S_j + pos_1, ... \\ v_N + \lambda_s \cdot S_j + pos_N\} \tag{2}$$

where x_{cls} represents the [cls] token. We extract global representations from vision transformers for person re-ID. To perform long-term matching, we supervise the network to focus on identity-specific features which are temporally invariant.

3.2. Attribute De-biasing Module

In the long-term re-ID scenario, it is logical for the feature representation to be agnostic of changes in clothes and related accessories. Therefore, it is essential to design a specific optimization objective that de-biases the final feature representation from features related to clothes or accessories. However, annotations associated with clothing information will most often be unavailable to perform this de-biasing.

In order to address this, we first use a network called Attribute Person Recognition (APR) [21], which is trained for person attribute recognition to predict a set of clothes and accessories from the input. Hence, the predicted attributes for an input image are given by AT = { pa_1, pa_2, pa_3, .., pa_K } where K is 27 and 23 attributes for the Market-

1501 [39] and the DukeMTMC-reID [40] datasets, respectively. The dashed line box shows several person attribute examples in Figure 2. We take these as the pseudo attribute ground truth labels for an input image. Given these pseudo labels, one way to enforce the feature extractors to learn a clothing-agnostic feature representation is to adopt a Gradient Reversal Layer (GRL) [5]. We incorporate this by adding an auxiliary module that consists of two sequential projection blocks, each having a linear layer followed by GeLU activation. Batch Normalization is done before the activation. The output of this auxiliary module is passed to a classification layer to predict the presence or absence of each attribute. Then, we reverse the gradients obtained from this module before passing them to the transformer-based feature extractor, as shown with a dashed arrow in Figure 2. Formally,

$$G_{feat} = \frac{\delta L_{attr}}{\delta \theta_{attr}} \cdot \frac{\delta \theta_{attr}}{\delta \theta_f} \\ RG_{feat} = -\lambda G_{feat} \tag{3}$$

where L_{attr} is the gradient associated with the auxiliary module parameterized by θ_{attr}. G_{feat} is a gradient of attribute prediction loss with respect to parameters θ_f associated with the transformer-based feature extractor. RG_{feat} refers to the revised gradient that is weighted by a factor λ.

3.3. Optimization

We optimize three objectives during training. As can be observed in Figure 2, features extracted from the transformer encoder are projected using linear transformations to get the visual identity embedding v^i_{ID}, which we use to optimize two losses (i) ID Loss and (ii) Triplet Loss following [10]. Identity objective $\mathcal{L}_{ID}$ is the cross-entropy loss without label smoothing. Mathematically,

$$\mathcal{L}_{ID} = -\sum_{i=1}^{n} y_i \cdot \log(x_i) \tag{4}$$

where y_i represents the ground truth identity, and x_i represents the predicted identity. Let $\mathcal{L}_T$ represent the triplet loss between an anchor sample a, positive sample p, and negative sample n. Formally,

$$\mathcal{L}_T = \log[1 + \exp(||v^a_{ID} - v^p_{ID}||^2_2 - ||v^a_{ID} - v^n_{ID}||^2_2)] \tag{5}$$

The third training objective is attribute prediction loss. The features extracted from the image encoder are projected separately using a different set of linear projections to get attribute embedding v^i_{attr}, which is used to predict attributes using a Binary Cross Entropy (BCE) loss.

$$\mathcal{L}_{attr} = -[y_i \cdot \log x_i + (1 - y_i) \cdot \log(1 - x_i)] \tag{6}$$

Here, x_i represents the predicted attribute label, and y_i represents the pseudo ground truth attribute labels which are the same as AT.

The final loss objective is the weighted summation of the three losses.

$$\mathcal{L} = \lambda_{ID} \cdot \mathcal{L}_{ID} + \lambda_T \cdot \mathcal{L}_T + \lambda_{attr} \cdot \mathcal{L}_{attr} \tag{7}$$

where λ_{ID}, λ_T, and λ_{attr} denote the weights of identity loss, triplet loss, and attribute loss, respectively. Here, the contribution from $\lambda_{attr} \cdot \mathcal{L}_{attr}$ is positive since the fully connected layers used for attribute predictions are required to be optimized to predict the correct attribute label. The gradient reversal, after the features are extracted, ensures that the image encoder learns to de-bias attribute features.

4. Experiments

4.1. Datasets

We evaluate the effectiveness and performance of our proposed model on two long-term cloth-changing re-ID datasets: LTCC [26] and NKUP [33]. The details of the datasets are described as follows.

LTCC [26] is an indoor clothes-changing re-ID dataset. It consists of 17,138 images of 152 identities wearing 478 outfits captured by 12 cameras for two months. On average, there are five clothing outfits for each person, with the number of outfit changes ranging from 2 to 14. Following [26], we split the LTCC dataset into training and testing sets. The training set in total consists of 77 identities. 46 out of these 77 identities are used to train the model in the cloth-changing scenario, whereas the remaining 31 subjects are used for training the model in the standard setting. Similarly, the test set contains 45 subjects used to evaluate the model's performance in the cloth-changing scenario and the remaining 30 identities are used for evaluation under the standard setting.

NKUP [33] is an indoor/outdoor clothes-changing re-ID dataset. It includes 9,738 images of 107 identities collected from 15 cameras for four months, 8 of which were installed in the outdoor environment. Among all the images, 5,336 images of 40 identities were used as the training set, while 332 and 4,070 images of 67 identities were used as the query and gallery images, respectively. The query set includes 3 to 10 images of each person, randomly selected from certain clothing styles. Then, the remaining images of the same person (with different clothing styles) were considered as the gallery images. Finally, images of individuals having only one clothing style were considered distractors and were added to the gallery samples. Mostly, subjects in the dataset appear in 2 or 3 different outfits.

Methods	LTCC				NKUP			
	Cloth-changing		Standard		Cloth-changing		Standard	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
LOMO [20] + KISSME [16]	10.8	5.3	26.6	9.1	-	-	-	-
LOMO [20] + XQDA [20]	11	5.6	25.4	9.5	-	-	-	-
PCB [30]	23.5	10	65.1	30.6	-	-	16.9	12.4
HACNN [19]	21.6	9.3	60.2	26.7	-	-	-	-
MuDeep [24]	23.5	10.2	61.9	27.5	-	-	-	-
RGA-SC [38]	31.4	14	65	27.5	-	-	-	-
MGN [32]	-	-	-	-	-	-	18.8	15
ISP [42]	27.8	11.9	66.3	29.6	-	-	-	-
Qian *et al.* [26]	26.2	12.4	71.4	34.3	-	-	-	-
GI-ReID [13]	28.1	13.2	73.6	36.1	-	-	-	-
FSAM [12]	38.5	16.2	73.2	35.4	-	-	-	-
LSD [35]	-	-	-	-	13.9	7.8	16.4	10.2
TransReID$_{base}$ [10]	67.0	29.6	93.8	82.5	21.8	14.4	24.0	18.2
AD-ViT (ours)	**72**	**34.2**	**94.8**	**84.3**	**23.6**	**16.9**	**27**	**18.9**

Table 1: Performance (%) comparison with the state-of-the-art methods. TransReID$_{base}$ [10] results are generated by us since the paper does not evaluate on the LTCC and NKUP datasets. 'Cloth-changing' and 'Standard' mean the cloth-changing setting and standard setting, respectively. The best performances are labeled in bold.

Backbones	Cloth-changing				Standard			
	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
ResNet50	22.5	61.7	73	79.1	62.1	86.5	92.7	93.8
ViT	**29.6**	**67**	**78.6**	**81.9**	**82.5**	**93.8**	**96.9**	**97.9**

Table 2: Performance (%) comparison of different backbones on the LTCC dataset. Note that they are basic backbone models and do not utilize side information and gradient reversal module.

4.2. Evaluation Protocols

To evaluate the person re-ID performance, we report the mean Average Precision (mAP) and Rank@k (R-k). Following the standard practice on LTCC [26] and NKUP [33], we report results on both cloth-changing and standard setting.

4.3. Implementation Details

Following [10], ViT was pre-trained on ImageNet and used as the backbone for our model. The input images are resized to 256×128. We split the image using the overlapping patch technique introduced in [10]. For data augmentation, we employ random horizontal flipping, padding, random cropping, and random erasing [41]. Each batch contains 64 images of 16 identities. SGD optimizer is employed with a momentum of 0.9 and a weight decay of 5e-4. The learning rate is initialized as $3.5 \cdot 10^3$ with cosine learning rate decay [23]. For the LTCC dataset, we set λ_{attr} = 1.0 and λ_{ID} = 1.0 in the cloth-changing setting and λ_{attr} = 2.0 and λ_{ID} = 0.8 in the standard setting. For the NKUP dataset, we set λ_{attr} = 1.0 and λ_{ID} = 1.0 in both scenarios. For all datasets, we set λ_s = 3.0, λ_T = 1.0.

4.4. Comparison with State-of-the-art Methods

Table 1 presents the performance of our proposed method on the long-term datasets: LTCC and NKUP. As can be observed, our method consistently outperforms the existing methods. On the LTCC dataset, compared to FSAM [12], the transformer-based backbone reproduced from TransReID (referred as TransReID$_{base}$) gave a substantial boost in performance. In the LTCC's cloth-changing scenario, AD-ViT achieves 5% improvement over TransReID$_{base}$ and 33.5% improvement over FSAM [12] for R-1. We also observe 18% improvement in mAP over FSAM [12]. In the NKUP dataset, one can note that the proposed method obtains 9.7% improvement in R-1 and 8.7% improvement in mAP compared to LSD [35]. These experiments show the suitability of the proposed model for the long-term cloth-changing scenario. Further, we can also note that using our proposed model, the performance under

Methods	SIE	GRL	LTCC				NKUP			
			Cloth-changing		Standard		Cloth-changing		Standard	
			R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Baseline			67	29.6	93.8	82.5	21.8	14.4	24	18.2
	✓		70	32.1	**94.8**	81.5	22.9	16.7	24	17.5
AD-ViT	✓	✓	**72**	**34.2**	**94.8**	**84.3**	**23.6**	**16.9**	**27**	**18.9**

Table 3: The ablation study of AD-ViT on the LTCC and NKUP datasets.

Methods	APR [21]		LTCC				NKUP			
			Cloth-changing		Standard		Cloth-changing		Standard	
	Models	Datasets	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Baseline			67	29.6	93.8	82.5	21.8	14.4	24	18.2
AD-ViT	ResNet50	Market1501	**72**	**34.2**	**94.8**	83.5	22.1	15.2	23	**19.6**
AD-ViT	ResNet50	DukeMTMC	68.3	32	92.7	81.5	**23.6**	**16.9**	**27**	18.9
AD-ViT	DenseNet121	Market1501	71	33.5	**94.8**	**84.3**	20.7	14.2	25	16.1
AD-ViT	DenseNet121	DukeMTMC	68.8	33.2	92.7	83.2	22.9	15.2	22	15.6

Table 4: Analysis of the performance of each GRL module using predicted attributes from different models. Market1501 and DukeMTMC denote the Market-1501 [39] and DukeMTMC-reID [40] datasets, respectively.

the standard setting is also improving, showing the robustness of our model.

4.5. Ablation Study of Different Backbones

In this section, we compare different types of feature extractor backbones: CNN-based and transformer-based. From Table 2, we can observe a significant performance gap between ResNet50 and ViT methods under both cloth-changing and standard scenarios. This difference can be attributed to the fact that the CNN-based model processes local neighborhoods of the input image individually and suffers from information loss on details caused by down-sampling operators. Considering the overall performances in both scenarios, we selected ViT model as a backbone for all our experiments.

4.6. Ablation Study of AD-ViT

We evaluate the benefits of employing SIE and GRL modules in AD-ViT in Table 3. In the cloth-changing setting, compared to the baseline, adding SIE module improves the performance by 3.0% for R-1 and 2.5% for mAP on LTCC. Similarly, for NKUP, we observe 1.1% and 2.3% improvement in R-1 and mAP, respectively. In the standard setting, compared to the baseline, SIE improves the performance by 1.0% on R-1 on LTCC. By combining SIE with GRL module in our proposed AD-ViT, we observe 5.0% improvement for R-1 under the cloth-changing setting for the LTCC. We also note that this performance improvement (1.8%) is also present in the NKUP dataset under similar settings. These experiments demonstrate the effectiveness

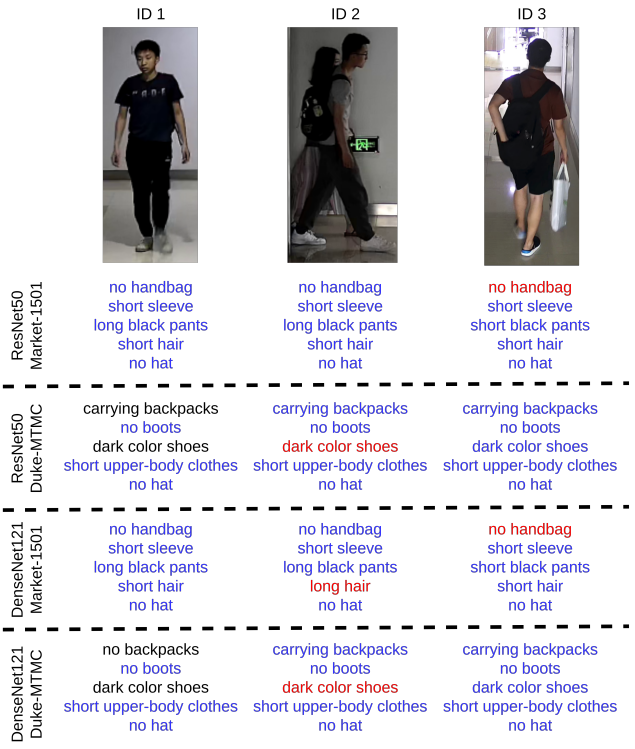

Figure 3: Examples of attribute predictions from different models and datasets. The images are from the LTCC dataset. The red, blue, and black descriptions represent incorrect, correct, and indistinguishable prediction labels, respectively.

of SIE and GRL in the proposed AD-ViT model.

4.7. Analysis of Different Attribute Predictions

We analyze the performance of our model using different Attribute Person Recognition (APR) networks. We use APR network [21] trained using two CNN backbones: ResNet50 and DenseNet121 on two different datasets: Market-1501 [39] and DukeMTMC-reID [40]. As indicated in Table 4, the performance of AD-ViT using predicted attributes from the ResNet50-based APR model trained on Market-1501 gives the best score for both R-1 and mAP in the LTCC's cloth-changing scenario. However, the performance of AD-ViT using predicted attributes from the ResNet50-based APR model trained on DukeMTMC-reID obtains the best results in NKUP's cloth-changing scenario. Therefore, we can observe a lot of performance variation across attribute prediction models and attribute datasets. This indicates that the predicted attribute labels are noisy. Likewise, the same holds true in Figure 3. Nevertheless, we observe that even with the noisy labels, de-biasing the ViT-based feature extractor helps improve the performance.

5. Conclusion

In this paper, we propose Attribute De-biased Vision Transformer (AD-ViT), which is a transformer-based attribute de-biasing architecture for long-term re-ID. The proposed method explored the creation of clothing and accessories agnostic feature representation using gradient reversal. Through rigorous experiments and ablation studies on long-term clothes-changing re-ID datasets, we demonstrate the value of our proposed method. Further, the performance improvement on the cloth-changing setting does not hamper the performance on the standard re-ID setting, indicating the robustness of the proposed model. Building on this, future works can explore adversarial domain adaptation techniques and novel transformer architectures to improve the performance further.

6. Acknowledgment

This material is based upon work partially supported by the National Science Foundation under Grant IIP #1822190.

References

[1] V. Bansal, G. L. Foresti, and N. Martinel. Cloth-changing person re-identification with self-attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 602–610, 2022. 2

[2] J. Berclaz, F. Fleuret, and P. Fua. Multi-camera tracking and atypical motion detection with behavioral maps. In *European conference on computer vision*, pages 112–125. Springer, 2008. 1

[3] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[4] 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. 3

[5] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 4

[6] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. 2015. 2

[7] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *Advances in neural information processing systems*, 31, 2018. 1, 2

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008. 1, 2

[9] L. He, J. Liang, H. Li, and Z. Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7073–7082, 2018. 1, 2

[10] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021. 3, 4, 5

[11] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European conference on computer vision*, pages 780–793. Springer, 2012. 1, 2

[12] P. Hong, T. Wu, A. Wu, X. Han, and W.-S. Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10513–10522, 2021. 2, 5

[13] X. Jin, T. He, K. Zheng, Z. Yin, X. Shen, Z. Huang, R. Feng, J. Huang, Z. Chen, and X.-S. Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14278–14287, 2022. 2, 5

[14] X. Jin, C. Lan, W. Zeng, and Z. Chen. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11165–11172, 2020. 1, 2

[15] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[16] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2288–2295. IEEE, 2012. 1, 2, 5

[17] K. W. Lee, N. Sankaran, D. Mohan, K. Davila, D. Fedorishin, S. Setlur, and V. Govindaraju. Bayesian personalized-wardrobe model (bp-wm) for long-term person re-identification. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2021. 2

[18] K. W. Lee, N. Sankaran, S. Setlur, N. Napp, and V. Govindaraju. Wardrobe model for long term re-identification and appearance prediction. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2018. 2

[19] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018. 5

[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 1, 2, 5

[21] 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. 3, 6, 7

[22] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 2

[23] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[24] X. Qian, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue. Leader-based multi-scale attention deep architecture for person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):371–385, 2019. 5

[25] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–667, 2018. 1, 2

[26] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 4, 5

[27] A. Rozantsev, M. Salzmann, and P. Fua. Beyond sharing weights for deep domain adaptation. *CoRR*, abs/1603.06432, 2016. 2

[28] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. *CoRR*, abs/1511.05547, 2015. 2

[29] X. Sun and L. Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 608–617, 2019. 1, 2

[30] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, pages 480–496, 2018. 2, 5

[31] F. Wan, Y. Wu, X. Qian, Y. Chen, and Y. Fu. When person re-identification meets changing clothes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 830–831, 2020. 2

[32] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282, 2018. 2, 5

[33] K. Wang, Z. Ma, S. Chen, J. Yang, K. Zhou, and T. Li. A benchmark for clothes variation in person re-identification. *International Journal of Intelligent Systems*, 35(12):1881–1898, 2020. 2, 4, 5

[34] X. Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013. 1

[35] E. Yaghoubi, D. Borza, B. Degardin, and H. Proença. You look so different! haven't i seen you a long time ago? *Image and Vision Computing*, 115:104288, 2021. 5

[36] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. Kweon. Pixel-level domain transfer. *CoRR*, abs/1603.07442, 2016. 2

[37] S.-I. Yu, Y. Yang, and A. Hauptmann. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3714–3720, 2013. 1

[38] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195, 2020. 5

[39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 4, 6, 7

[40] 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, 6, 7

[41] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 5

[42] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang. Identity-guided human semantic parsing for person re-identification. In *European Conference on Computer Vision*, pages 346–363. Springer, 2020. 5