Rethinking Person Re-Identification via Semantic-based Pretraining

SUNCHENG Xiang and DAHONG QIAN, School of Biomedical Engineering, Shanghai Jiao Tong University, China
JINGSHENG GAO, ZIRUI ZHANG, TING LIU, and YUZHUO FU, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China

Pretraining is a dominant paradigm in computer vision. Generally, supervised ImageNet pretraining is commonly used to initialize the backbones of person re-identification (Re-ID) models. However, recent works show a surprising result that CNN-based pretraining on ImageNet has limited impacts on Re-ID system due to the large domain gap between ImageNet and person Re-ID data. To seek an alternative to traditional pretraining, here we investigate semantic-based pretraining as another method to utilize additional textual data against ImageNet pretraining. Specifically, we manually construct a diversified FineGPR-C caption dataset for the first time on person Re-ID events. Based on it, a pure semantic-based pretraining approach named VTBR is proposed to adopt dense captions to learn visual representations with fewer images. We train convolutional neural networks from scratch on the captions of FineGPR-C dataset, and then transfer them to downstream Re-ID tasks. Comprehensive experiments conducted on benchmark datasets show that our VTBR can achieve competitive performance compared with ImageNet pretraining—despite using up to 1.4x fewer images, revealing its potential in Re-ID pretraining. Our source code is also publicly available at https://github.com/JeremyXSC/VTBR.

CCS Concepts: • Computing methodologies → Object identification; Feature selection; Image representations;

Additional Key Words and Phrases: Person re-identification, synthetic data, efficient training

ACM Reference format:
Suncheng Xiang, Dahong Qian, Jingsheng Gao, Zirui Zhang, Ting Liu, and Yuzhuo Fu. 2023. Rethinking Person Re-Identification via Semantic-based Pretraining. ACM Trans. Multimedia Comput. Commun. Appl. 20, 3, Article 90 (December 2023), 17 pages.
https://doi.org/10.1145/3628452
1 INTRODUCTION

CNN-based pretraining on ImageNet is a dominant paradigm in computer vision. As many vision tasks are related, it is expected a deep learning model, pretrained on one dataset, to help another downstream task. It is now common practice to pretrain the backbones of object detection [10] and segmentation [19] on ImageNet [3] dataset. In the field of person Re-ID, most of works [1, 8, 42, 43, 51] try to leverage models pretrained on ImageNet to mitigate the shortage of person Re-ID data, which has achieved remarkable performance. However, this practice has been recently challenged by [7], who show a surprising result that such ImageNet pretraining may not be the best choice for the re-identification task due to the intrinsic domain gap between ImageNet and person Re-ID data. Additionally, some research works [5, 31] also indicate that learning visual representations from textual annotations can be more competitive to methods based on ImageNet pretraining, which has attracted considerable attention from both academia and industry worldwide. For these reasons, there has been increasing interest for us to explore novel vision-and-language pretraining strategies which can replace the traditional ImageNet-based pretraining paradigm on Re-ID tasks. Unfortunately, existing datasets [15, 32, 44, 45, 47, 54, 57] in Re-ID community are all of limited scale due to the costly efforts required for data collection and annotation, especially since none of them has diversified attributes to obtain dense captions for pedestrians, which fails to satisfy the need of semantic-based pretraining in Re-ID task.

Targeting to address above mentioned limitations, we start from two aspects, namely, data and methodology. From the data perspective, we construct a FineGPR-C caption dataset for the first time on person Re-ID events, which involves human describing event in a fine-grained manner. From the methodology perspective, we propose a pure VirTex Based Re-ID pretraining approach named VTBR, which uses transformers to learn visual representations from textual annotations, the overview of our framework is illustrated in Figure 1. Particularly, we jointly train a CNN-based network and transformer-based network from scratch using image caption pairs for the task of image captioning. Then, we transfer the learned residual network to downstream Re-ID tasks. In general, our method seeks a common vision-language feature space with discriminative learning constraints for better practical deployment.

The initial motivation for this research comes from a comprehensive study of Re-ID pretraining. In the course of our efforts, we noticed that semantic captions can provide a denser learning signal than traditional unsupervised or supervised learning [5], so using language supervision on Re-ID task is more appealing, which can provide supervision for learning transferable visual representations with better data-efficiency than other previous approaches. Another benefit of textual annotation is simplified data collection. Traditional labeling procedure of real pedestrian data always costs intensive human labor, sometimes even involving person privacy concerns and data security problems, which brings researcher a serious challenge for dataset collection. In contrast, natural language description from fine-grained attributes on synthetic data do not require an explicit category and can be easily labeled by non-expert workers, leading to a simplified data labeling procedure without ethical issues regarding privacy. To the best of our knowledge, we are among the first attempts to use textual features to perform pretraining for downstream Re-ID tasks. We hope this study and FineGPR-C caption dataset will serve as a solid baseline in semantic-based Re-ID pretraining and pave a path for the community to move forward.

As a consequence, the major contributions of our work can be summarized into three-fold:

- We manually construct a FineGPR-C caption dataset for the first time to enable the semantic pretraining for Re-ID task.
- Based on it, a semantic-based pre-training approach named VTBR is proposed to learn visual representations from textual annotations on Re-ID event.
Fig. 1. The overview of our Re-ID pretraining framework. First, we jointly train ResNet and Transformer using image caption pairs for the task of image captioning. Then, we transfer the learned ResNet as the backbone of the downstream Re-ID task.

— Comprehensive experiments show that our VTBR method matches or exceeds the performance of existing methods for supervised or unsupervised pretraining on ImageNet with fewer images, which reveals the applicability of semantic-based pretraining with new insights.

In the rest of the article, we first review some related works of person re-identification methods and previous pre-training method in Section 2. Then in Section 3, we give more details about the first FineGPR-C caption dataset, as well as learning procedure of the proposed VTBR pretraining method. Extensive evaluations compared with state-of-the-art methods and comprehensive analyses of the proposed approach are elaborated in Section 4. Conclusion and Future Works are given in Section 5.

2 RELATED WORKS
In this section, we have a brief review of the related work of person Re-ID method and pretraining approach. The mainstream idea of the existing methods is to learn a robust model for feature representation.

2.1 Person Re-ID Methods
Actually, there are mainly two kinds of feature learning paradigms for person Re-ID tasks: (1) Handcrafted-based method and (2) Deep learning-based approach, which are introduced as follows:

Traditional research works [6, 50, 53] related to hand-crafted systems for person re-ID aim to design or learn discriminative representation or pedestrian features. For example, Farenzena et al. [6] propose an appearance-based method for these situations where the number of candidates varies...
continuously. Zhao et al. [53] design a novel approach of learning mid-level filters from automatically discovered patch clusters for person re-identification. Besides directly using mid-level color and texture features, some methods [50] also strive to learn a similarity metric from image pixels directly, which can jointly learn the color feature, texture feature, and metric in a unified manner. Unfortunately, these hand-crafted feature based models always fail to produce competitive results on large-scale datasets. The main reason is that these early works are mostly based on heuristic design, and thus, they could not learn optimal discriminative features on current large-scale dataset.

Recently, benefited from the advances of deep neural networks and availability of large-scale datasets, person Re-ID performance in supervised learning has been significantly boosted to a new level [41, 44], e.g., Xiang et al. [41] propose a feature fusion strategy based on traditional convolutional neural network with attention mechanism, which learns robust feature extraction and reliable metric learning in an end-to-end manner. Lopez et al. [20] propose an end-to-end multi-modal CNN that combines image and context information with an attention module for scene recognition. Ning et al. [25] design a feature selection network that combines global and local fine-grained features for person Re-ID task. There is also a study [24] that locates all the valuable areas of the features on the basis of the joint weak saliency mechanism and attention-aware model. Besides, some recent works [4, 39] attempt to address unsupervised domain adaptation based on Generative Adversarial Network (GAN) model. However, these approaches always require abundant computing resources to achieve satisfactory performance, and leveraging GAN network is unable to guarantee the quality of generated images. Additionally, previous methods [2, 26, 27] either focus on designing various deep CNN structures to learn discriminative feature embeddings or strive to explore better loss functions for deep neural network training [12, 33, 36, 48]. In essence, these works always leverage models pretrained on ImageNet dataset to mitigate the shortage of person Re-ID data, which suffers from the limitation of large domain gap between ImageNet and person Re-ID data.

2.2 Person Re-ID Pretraining

ImageNet pretraining is a dominant paradigm in vision community. Recently, some researches have shown a surprising result that CNN-based pretraining on ImageNet has limited impacts on Re-ID system due to the large domain gap between ImageNet and person Re-ID data. On the one hand, benefit from large-scale pedestrian datasets, researchers try to leverage large-scale Re-ID datasets to perform pretraining, which can help to learn a discriminative feature representations with high quality. Specifically, Fu et al. [7] propose an unsupervised pretraining strategy with contrastive learning based on LUPerson dataset. Yang et al. [49] design an unsupervised pre-training framework for Re-ID based on the contrastive learning. Besides, Luo et al. [22] propose a self-supervised pre-training strategy for transformer-based person re-identification task, which can further reduce the domain gap and accelerate the pre-training. On the other hand, to mine the potential of semantic features, several studies [5, 31] try to learn visual representations from textual annotations on basic computer vision tasks, e.g., Radford et al. [31] demonstrate that learning visual models from natural language supervision is an efficient way to learn the state-of-the-art image representation, which has attracted considerable attention from both the academia and industry worldwide. Since semantic feature is critical for representation learning, many research works [13, 18, 40] try to leverage the semantic feature to boost the performance of person Re-ID task, for example, Lin et al. [18] build an attribute-person recognition network to exploit both identity labels and attribute annotations for better semantic feature representation on person Re-ID task. Jeong et al. [13] present an efficient and effective framework with semantic affinities for attribute-based person search. Xiang et al. [40] propose a deep multimodal representation
learning network to elaborate rich semantic knowledge for assisting in representation learning during the pre-training. Unfortunately, these methods fail to achieve satisfactory performance due to the limited scale of textural annotations on person Re-ID event. Inspired from these studies, in this article, we manually construct a FineGPR-C caption dataset for person Re-ID events, based on it, a pure semantic-based Re-ID pretraining framework named VTBR is proposed to seek a common vision-language feature space with discriminative learning constraints for robust representation and better practical deployment.

Although VTBR inherits the structure of previous transformer-based CLIP [31] and VirTex [5], there exists some significant new designs in VTBR to allow it work for a very different manner: First, we only employ standard pretraining strategy in a supervised fashion with labeled image pairs, while CLIP adopts contrastive self-supervised learning with unlabeled image pairs from COCO dataset [17], and VirTex leverages the semantic density of captions to learn visual representations for downstream detection tasks, note that VirTex employs the dense caption of COCO Captions dataset (e.g., five captions per image) to increase the image-caption pairs by five-fold, leading an expensive annotation costs in terms of worker hours, which indicate these works are more complex; Second, during the testing stage, both Text Encoder and Image Encoder are adopted for classification in previous CLIP, while our VTBR method only employs single visual backbone for downstream Re-ID task, allowing our method to be more flexible in real-world scenarios. Finally, this is the first time as far as we know, to significantly explore the potential of semantic features on Re-ID pretraining, from which we have proved that learning visual representation using textual annotations can be competitive to methods based on both supervised and unsupervised learning on ImageNet. We also hope that our dataset and method will shed light on some related researches to move forward, especially for semantic-based pretraining.

3 PROPOSED METHOD

3.1 Problem Formulation

For the pretraining of Re-ID task, given a labeled source dataset $S = \{x_1, x_2, \ldots, x_{N_s}\}$, consisting of $N_s$ person images with manually annotated labels $Y = \{y_1, y_2, \ldots, y_{N_s}\}$, our goal is to learn a pretrained embedding function $\phi(\theta)$ from labeled source dataset $S$. We also have an unlabeled target dataset $T = \{t_1, t_2, \ldots, t_M\}$. Note that there is non-overlapping in terms of identity between source domain and target domain in open set domain adaptation. Finally, by leveraging both labeled source images from $S$ and unlabeled target samples in $T$, we can train a discriminative Re-ID model that generalized well in supervised or domain adaptive Re-ID task.

3.2 FineGPR-C Caption Dataset

Data is the life-blood of training deep neural network models and ensuring their success. For the person Re-ID task, sufficient and high-quality data are necessary for increasing the model’s generalization capability. In this work, we ask the question: can we construct a person dataset with captions which can be used as semantic-based pretraining on Re-ID task? To answer this question, we revisit the previously developed FineGPR [46] dataset, which contains fine-grained attributes such as viewpoint, weather, illumination, and background, as well as 13 accurate annotations at the identity level. More details about the attribute distribution of pedestrian is available in Figure 2.

To provide data foundation for semantic-based pretraining, on the basis of FineGPR, we introduce a dynamic strategy to generate high-quality captions with fine-grained attribute annotations for semantic-based pretraining. To be more specific, we rearrange the different attributes as word embeddings into caption expressions at the different position and then generate semantically dense caption containing high-quality description, this gives rise to our newly-constructed FineGPR-C
Fig. 2. Attributes illustration at the identity level. We manually annotate these attributes at the identity level in a fine-grained manner.

Fig. 3. Some exemplars of semantic caption in our FineGPR-C caption dataset, which is generated using our dynamic caption generating strategy.

caption dataset. Some exemplars of FineGPR-C dataset are depicted in Figure 3. It is worth mentioning that different pedestrian images have different captions by the different regular expressions.

During the caption generation process, we found that there exists serious redundancy among the different attributes in FineGPR-C, especially for some attribute that appears with a larger probability. As shown in Figure 4, for example, the attribute of adult in terms of Age accounts for more than three-quarters (nearly 1,049/1,150) of total identities in FineGPR-C caption dataset, which may trigger a new problem that pedestrian attributes with high probability will degrade the diversity of generated caption dataset. To address this problem, we introduce a **Refined Selecting (RS)** strategy to increase the inter-class diversity of different identities and minimize the intra-class variation of the same identity. Particularly, we set a threshold $\lambda$ to control the appearing...
Rethinking Person Re-Identification via Semantic-based Pretraining

Fig. 4. Distribution of attributes in terms of identity for FineGPR-C dataset. Please zoom in for the best view.

Fig. 4. Distribution of attributes in terms of identity for FineGPR-C dataset. Please zoom in for the best view.

probability of attributes in the final caption sentence $c$, which can dynamically select some representative pedestrian attributes that appear with a lower probability, finally, the formula can be expressed as

$$c = \{w_1, a_1, w_2, a_2, \ldots, w_K, a_K\}, \text{ if } P_{a_1}, P_{a_2}, \ldots, P_{a_K} \leq \lambda,$$

(1)

where $K$ indicates the total number of identities, $w_1, w_2, \ldots, w_K$ denote fixed sentence words, and $a_1, a_2, \ldots, a_K$ indicate labeled pedestrian attributes, respectively. $P_{a_1}, P_{a_2}, \ldots, P_{a_K}$ represents the appearing probability of the corresponding attribute annotation $a_1, w_2, \ldots, a_K$ in FineGPR, respectively. For example, if an attribute of a pedestrian appears with a larger probability ($\geq \lambda$), this pedestrian attribute will not be selected into the final caption sentence $c$. In contrast, this attribute can be dynamically selected into the final caption sentence $c$, if an attribute of a pedestrian appears with a small probability ($\leq \lambda$). The main purpose of this strategy is to reduce the serious redundancy among the different attributes in FineGPR-C dataset, thus, increasing the inter-class diversity of different identities and minimizing the intra-class variation of the same identity. In principle, our overall goal for constructing FineGPR-C caption dataset is to improve the caption’s discriminative ability according to their attribute distribution, so the generated caption (token by token) will be more diversified and contain more discriminative information.

To this end, even though newly-built FineGPR-C dataset is based on the previous FineGPR dataset, FineGPR-C is actually different from FineGPR since it is constructed with fine-grained pedestrian attributes by our RS strategy, which lays a data foundation for semantic pretraining framework VTBR. Please note that FineGPR-C is also the first caption dataset for person Re-ID events, which will serve as a solid baseline in semantic pretraining and can greatly advance the research in Re-ID community.

3.3 Proposed VTBR Approach

In order to learn deep visual representations from textual annotations for Re-ID task, we introduce a semantic-based pretraining method VTBR based on our newly-built FineGPR-C dataset. As illustrated in Figure 5, our VTBR framework consists of a visual backbone ResNet-50 [11] and semantic backbone Transformer [37], which extracts visual features of images and textual features of caption, respectively. Firstly, the visual features extracted from ResNet-50 are used to predict captions of pedestrian images by transformer networks. Following the [5], we use projection layer to receive features from the visual backbone, then put them to the textual head to predict captions with transformers for images, which adopts multiheaded self-attention both to propagate...
Fig. 5. The framework of our vision-language supervised pretraining approach VTBR, which consists of a visual backbone ResNet-50 and Transformer. The visual backbone extracts visual features, and transformers predict captions via bidirectional language modeling on the basis of visual features. After pretraining, the visual backbone is transferred to downstream Re-ID tasks.

information of caption tokens and then provide a learning signal to the visual backbone during pretraining. To be more specific, the semantic backbone (Textual Head) comprises two identical language models that predict captions in forward and backward directions, respectively, which use mult headed self-attention both to propagate information along the sequence of caption tokens, as well as to fuse visual and textual features. During training, the Forward Transformer Decoder receives image features from the visual backbone, and a dense semantic caption in our FineGPR-C dataset. Image features are a matrix of shape \(7 \times 7 \times H\) giving an \(H\)-dimensional vector for each of the \(7 \times 7\) positions in the final layer of the visual backbone. Note that this projection layer is not used in downstream tasks. During training, we use the log-likelihood loss function to train the visual and semantic backbones in an end-to-end manner, which can be written as

\[
L = \sum_{k=1}^{K+1} \log \left( p \left( T, V; \psi_f, \phi \right) \right) + \sum_{k=0}^{K} \log \left( p \left( T, V; \psi_b, \phi \right) \right),
\]

where \(\psi_f, \psi_b\) and \(\phi\) mean forward transformer, backward transformer and ResNet-50, respectively. \(T\) and \(V\) denote textual features and visual features separately. Log-probabilities are predicted by the linear layer of the last Transformer layer over the token vocabulary. It is worth mentioning that both visual and semantic backbones are jointly trained to maximize the log-likelihood of the correct caption tokens. Instead of adopting pretrained weight on ImageNet dataset, we train our entire VTBR model from scratch on our FineGPR-C caption dataset, whereas they rely on pretrained transformer to extract textual features.

After obtaining the pretrained model based on our FineGPR-C caption dataset, we perform downstream Re-ID evaluation\(^1\) continuously. Specifically, we adopt global features extracted by visual backbone ResNet-50 to perform metric learning. Figure 6 gives a detailed illustration of the

\(^1\)In this work, we adopt a widely used open-source Re-ID backbone in [21], more details can be available at https://github.com/michuanhaohao/reid-strong-baseline
Fig. 6. The illustration of training procedure for Re-ID tasks. First, the deep features of input images are extracted by CNN-based network, then, two commonly used loss functions (e.g., Cross-entropy loss and Triplet loss) are adopted for deep metric learning. Finally, the fine-tuned model is employed for Re-ID evaluation.

Training procedure for Re-ID tasks. During the training stage, the visual backbone ResNet-50 is trained with commonly used softmax cross-entropy loss and triplet loss on the basis of the FineGPR and FineGPR-C caption dataset. Then, we transfer the learned CNN model to the downstream Re-ID task. During the period of inference, the deep features of input images extracted by CNN-based network are used to perform metric learning. It is worth mentioning that we only modify the output dimension of the latest fully-connected layer to the number of training identities [45], and we extract the 2,048-dim pool-5 vector for retrieval under the Euclidean distance. Finally, we can obtain the top-n ranking list as the model output for some query images on the standard Re-ID datasets.

4 EXPERIMENTAL RESULTS

4.1 Datasets

In this article, we conduct experiments on three large-scale public datasets, which include Market-1501 [54], DukeMTMC-reID [32, 57], and CUHK03 [15] datasets.

Market-1501 [54] contains 32,668 labeled images of 1,501 identities captured from campus in Tsinghua University. Each identity is captured by at most six cameras. The training set contains 12,936 images from 751 identities and the test set contains 19,732 images from 750 identities.

DukeMTMC-reID [32, 57] is collected from Duke University with eight cameras, it has 36,411 labeled images belonging to 1,404 identities and contains 16,522 training images from 702 identities, 2,228 query images from another 702 identities and 17,661 gallery images.

CUHK03 [15] contains 14,097 images of 1,467 identities. Following the CUHK03-NP protocol [58], it is divided into 7,365 images of 767 identities as the training set, and the remaining 6,732 images of 700 identities as the testing set.

In our experiments, we follow the standard evaluation protocol [54] used in Re-ID task, and adopt mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) at rank-1 and rank-5 for performance evaluation on downstream re-ID task.
4.2 Implementation Details

For the pretraining of VTBR, we apply standard random cropping and normalization as data augmentation. Following the training procedure in [5], we adopt SGD with momentum 0.9 and weight decay $10^{-4}$ wrapped in LookAhead [14] with $\alpha = 0.5$ and 5 steps. We empirically set the $\lambda = 0.8$ in Equation (1). The max learning rate of visual backbone is $2 \times 10^{-1}$; learning rate of the textual head is set as $1 \times 10^{-3}$. For the downstream Re-ID task, we closely follow a widely used open-source project [21] as standard baseline, which is built only with commonly used softmax cross-entropy loss [52] and triplet loss [12] on vanilla ResNet-50 with no bells and whistles. Following the practice in [21], the batch size of training samples is set as 64. As for triplet selection, we randomly selected 16 persons and sampled four images for each identity, $m$ is set as 0.5 as triplet margin. Adam method and warmup learning strategy are also adopted to optimize the model. All the experiments are performed on PyTorch [28] with two Nvidia GeForce RTX 3090 GPUs on a server equipped with an Intel Xeon Gold 6240 CPU.

4.3 Important Parameter

In this section, we evaluate the impacts of parameter $\lambda$ in Equation (1), which controls the probability of attributes appears in the final caption sentence $c$. As depicted in Figure 7, it can be easily observed that when $\lambda$ is small, the performance is not optimal because the selected attribute is way too limited to a very small portion, and thus our VTBR could not mine the relevant image regions or discriminative parts to learn semantic features. The $\lambda$ should also not be set too large, otherwise, the performance drop dramatically since generated caption can not maintain a high diversity. Specially, $\lambda = 0.8$ yields the best accuracy for Re-ID task.

4.4 Supervised Fine-tuning

In this work, the caption data for Re-ID event is the fundamental part of the semantic-based pretraining baseline. Here, we adopt supervised fine-tuning performance on real datasets as the indicator to show the quality of FineGPR-C caption dataset. From Table 1, we can obviously observe that the results of supervised learning are significantly promoted by using our method. For example, when training and testing on Market-1501 with ImageNet pretrained model, we can only achieve a rank-1 accuracy of $94.3\%$, while our VTBR method on FineGPR-C can obtain a
Table 1. Comparisons between Traditional CNN-based and Semantic-based VTBR Pretraining on Supervised Re-ID Tasks

| Pretrain             | S   | #Imgs | Market-1501 | DukeMTMC | CUHK03 |
|----------------------|-----|-------|-------------|----------|--------|
|                      |     |       | Rank-1 mAP  | Rank-1 mAP | Rank-1 mAP |
| ResNet (ImageNet)    | ×   | 1.28 M| 94.3        | 85.0     | 86.7   |
| ResNet (FineGPR)     | ×   | 2.00 M| 85.5        | 74.2     | 63.9   |
| VTBR (FineGPR-C)     | ✓   | 1.83 M| 93.6        | 83.7     | 85.0   |
| VTBR+RS (FineGPR-C)  | ✓   | 0.91 M| 94.9        | 85.3     | 87.3   |

“S” denotes semantic feature. For different pretraining methods, “ResNet (ImageNet)” means we pretrain ResNet-50 on ImageNet dataset, same for ResNet-50 (FineGPR), VTBR (FineGPR-C), and VTBR+RS (FineGPR-C). Orange indicates the best and Blue the second best.

Table 2. Comparisons between Traditional CNN-based and Semantic-based VTBR Pretraining on Domain Adaptive Re-ID Tasks

| Pretrain             | S   | #Imgs | DukeMTMC→Market-1501 | Market-1501→DukeMTMC |
|----------------------|-----|-------|-----------------------|-----------------------|
|                      |     |       | Rank-1 Rank-5 mAP     | Rank-1 Rank-5 mAP     |
| ResNet (ImageNet)    | ×   | 1.28 M| 48.0 64.1 21.7        | 24.5 38.8 13.8        |
| ResNet (FineGPR)     | ×   | 2.00 M| 44.2 62.8 20.5        | 20.8 33.5 10.2        |
| VTBR (FineGPR-C)     | ✓   | 1.83 M| 45.9 64.8 21.2        | 21.3 34.6 10.9        |
| VTBR+RS (FineGPR-C)  | ✓   | 0.91 M| 50.6 67.7 23.8        | 24.3 38.4 13.5        |

“S” denotes semantic feature. For different pretraining methods, “ResNet (ImageNet)” means we pretrain ResNet-50 on ImageNet dataset, same for ResNet-50 (FineGPR), VTBR (FineGPR-C), and VTBR+RS (FineGPR-C). Orange indicates the best and Blue the second best.

competitive performance of 93.6%. After employing the Refined Selecting strategy, our VTBR+RS reaches a remarkable performance of 94.9% with 1.4× fewer pretraining images (0.91 M vs. 1.28 M), leading to a record mAP performance of 85.3%. Not surprisingly, same performance gain can also be achieved on DukeMTMC-reID dataset. The success of VTBR can be largely contributed to the discriminative features learned by semantic captions in a data-efficient manner.

4.5 Unsupervised Domain Adaption

Our semantic-based pretraining method enjoys the benefits of flexible corner scenarios of domain adaptive Re-ID tasks, where labeled data in target domain is hard to obtain. In this section, we present four domain adaptive Re-ID tasks on several benchmark datasets. More detailed results can be seen in Tables 2–4. For instance, when trained on DukeMTMC-reID dataset, it can be easily observed that our VTBR+RS achieves a significant rank-1 performance of 50.6% and 5.7% on Market-1501 and CUHK03, respectively, outperforming the ImageNet pretraining by +2.6% and +0.8% in terms of rank-1 accuracy. When trained on Market-1501 dataset, our method can also lead to an obvious improvement of +1.9% on CUHK03 in rank-1 accuracy. However, when tested on DukeMTMC-reID dataset, it was surprising to find that our method obtained a slightly inferior performance than ImageNet pretraining (mAP 13.5% vs. 13.8%, 0.91 M vs. 1.28 M images). We suspect that captions generated on FineGPR have an obvious domain gap with DukeMTMC-reID dataset since there are some occlusions and multiple persons in the queries, which will undoubtedly degrade the performance of our method.

4.6 Comparison with Other Methods

In this section, we compare our results with existing methods of CNN architecture in Table 5. Note that we do not apply any post-processing method like Re-Rank [58] in our approach. As we can see,
Table 3. Performance Comparison with Other Baselines of CNN Architecture on Unsupervised Domain Adaptation Re-ID Tasks

| Methods                  | DukeMTMC→Market-1501 | Market-1501→DukeMTMC |
|--------------------------|-----------------------|----------------------|
|                          | Rank-1    | Rank-5 | mAP | Rank-1    | Rank-5 | mAP |
| LOMO+XQDA [16]           | 27.2      | 41.6   | 8.0 | 12.3      | 21.3   | 4.8 |
| BoW+XQDA [54]            | 35.8      | 52.4   | 14.8| 17.1      | 28.8   | 8.3 |
| UMDL [29]                | 34.5      | 52.6   | 12.4| 18.5      | 31.4   | 7.3 |
| PTGAN [39]               | 38.6      | 57.3   | 15.7| 27.4      | 43.6   | 13.5|
| VTBR+RS (Ours)           | 50.6      | 67.7   | 23.8| 24.3      | 38.4   | 13.5|

*Orange* indicates the best and *Blue* the second best.

Table 4. Comparisons between Traditional CNN-based and Semantic-based VTBR Pretraining on Domain Adaptive Re-ID Tasks

| Domain Adaptation → | DukeMTMC→CUHK03 | Market→CUHK03 |
|---------------------|-----------------|---------------|
| Pretrain ↓          | Rank-1 | Rank-5 | mAP | Rank-1 | Rank-5 | mAP |
| ResNet (ImageNet)   | ×      | 1.28 M | 4.9 | 11.6   | 5.6 | 3.9 | 8.6 | 4.0 |
| ResNet (FineGPR)    | ×      | 2.00 M | 4.7 | 11.3   | 5.5 | 4.5 | 11.5 | 4.3 |
| VTBR (FineGPR-C)    | ✓      | 1.83 M | 4.9 | 11.4   | 5.2 | 5.4 | 11.8 | 5.8 |
| VTBR+RS (FineGPR-C) | ✓      | 0.91 M | 5.7 | 13.1   | 5.7 | 5.8 | 13.2 | 6.2 |

*S* denotes semantic feature. For different pretraining methods, “ResNet (ImageNet)” means we pretrain ResNet-50 on ImageNet dataset, same for ResNet-50 (FineGPR), VTBR (FineGPR-C), and VTBR+RS (FineGPR-C). *Orange* indicates the best and *Blue* the second best.

Table 5. Performance Comparison with Other Baselines of CNN Architecture on Supervised Re-ID Tasks

| Methods   | Market-1501 | DukeMTMC | CUHK03 |
|-----------|-------------|----------|--------|
|           | Rank-1     | mAP      | Rank-1 | mAP  | Rank-1 | mAP  |
| BoW+XQDA  [54] | 44.4  | 20.8     | 25.1   | 12.2 | 6.4    | 6.4  |
| LOMO+XQDA [16] | 43.8  | 22.2     | 30.8   | 17.0 | 12.8   | 11.5 |
| SVDNet    [34] | 82.3  | 62.1     | 76.7   | 56.8 | 41.5   | 37.2 |
| CASN(IDE) * [55] | 92.0  | 78.0     | 84.5   | 67.0 | 57.4   | 50.7 |
| DaRe [38]  | 86.4      | 69.3   | 75.2   | 57.3 | 55.1   | 51.3 |
| FD-GAN    [9]  | 90.5      | 77.7   | 80.0   | 64.5 | 92.6   | 91.3 |
| PN-GAN    [30] | 92.93 | 80.19    | 73.58  | 53.20 | 79.76  | –   |
| PCB+RPP*  [35] | 93.8  | 81.6     | 83.3   | 69.2 | 63.7   | 57.5 |
| DG-Net    [56] | 94.8   | 86.0     | 86.6   | 74.8 | –      | –    |
| VTBR+RS (Ours) | 94.9  | 85.3     | 87.3   | 76.8 | 61.9   | 59.3 |

* represents the attention-based method. *Orange* indicates the best and *Blue* the second best.

we can achieve state-of-the-art performance on Market-1501 and DukeMTMC-reID dataset with considerable advantages, by simply applying our semantic-based pre-training strategy, we can obtain a remarkable mAP performance of 85.3% and 76.8% on Market-1501 and DukeMTMC dataset, respectively, leading a significant Rank-1 improvement of +0.1% and +0.7% when compared with second best method DG-Net [56]. Surprisingly, we also find an interesting phenomenon that performance of VTBR+RS is slightly inferior and less competitive compared with FD-GAN [9] on CUHK03 dataset (Rank-1 92.6% vs. 61.9%). It is probably because that the attention or within-part
consistency among image of same pedestrian in CUHK03 dataset can produce principled supervisory signals to baseline CNN architecture, which can greatly enhance the pedestrian discriminability of attention-based model on Re-ID task.

4.7 Visualization

In order to verify the effectiveness of our proposed VTBR method, we show more qualitative examples of Grad-CAM [23] visualizations in Figure 8. In fact, the Grad-CAM is a package with some methods for Explainable AI for computer vision. This can be used for diagnosing model predictions, either in production or while developing models. To be more specific, given an image and a class of interest as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. This signal is then backpropagated to the rectified convolutional feature maps of interest, which we combine to compute the coarse Grad-CAM localization which represents where the model has to look to make the particular decision. Finally, we pointwise multiply the heatmap with guided backpropagation to get Guided Grad-CAM visualizations which are both high-resolution and concept-specific. Compared with ImageNet pretraining method, we observe that our model attends to relevant image regions or discriminative parts for making caption predictions, indicating that VTBR can greatly help the model learn more global context information and meaningful visual features with better

Fig. 8. Visualization of attention maps with Grad-CAM [23]. (a) Original images, (b) CNN-based pretraining method on ImageNet, (c) Our semantic-based VTBR method on FineGPR-C caption dataset. It can be easily observed that semantic pretraining method can capture global context information and more discriminative parts, which are further enhanced in our proposed VTBR method for better performance.
Table 6. Complexity Analysis of Our Proposed Method in Terms of FLOPs, Memory Cost, and Parameters, Respectively

| Methods            | DukeMTMC→Market-1501 | Market-1501→DukeMTMC |
|--------------------|-----------------------|-----------------------|
|                    | #FLOPs    | Memory | #Params  | #FLOPs    | Memory | #Params  |
| VTBR (Ours)        | 60.0 G    | 24 G×2 | 285.96 M | 60.0 G    | 24 G×2 | 287.10 M |

semantic understanding, which significantly makes our semantic-based pretraining VTBR model more robust to perturbations.

4.8 Discussion

According to the experiments from Table 1, Tables 2 and 4, semantic-based pretraining strategy demonstrates its competitiveness and priority since it can bridge the gap between ImageNet and downstream person Re-ID data. Additionally, a complexity analysis of the proposed method in terms of FLOPs, Memory Cost, and Parameters is also provided in the Table 6. Despite its promising performance on person Re-ID, we note that there are several limitations in our VTBR method.

First, the premise of our semantic-based VTBR method is that the caption dataset for downstream task is required to be constructed based on domain-specific knowledge and have a high diversity. However, there exists a fact that different vision tasks have different caption paradigm, this means the caption generation process may be repeated by carefully designed strategy when downstream task makes a change in different scenarios;

Second, the generated caption dataset FineGPR-C may still have domain gap with some collected datasets (e.g., DukeMTMC-reID) in terms of viewpoint, weather, illumination, as well as background, which may bring negative impacts and degrade the performance for downstream Re-ID task to a certain extent. We believe addressing these challenges are promising direction of our work for future research.

5 CONCLUSION AND FUTURE WORK

This article takes a big step forward to rethink person re-identification via semantic-based pretraining. Specially, we construct the first FineGPR-C caption dataset for person Re-ID events, which covers human describing in a fine-grained manner. Based on it, we present a simple yet effective semantic-based pretraining method to replace the ImageNet pretraining, which helps to learn visual representations from textual annotations on downstream Re-ID task. Extensive experiments conducted on several benchmarks show that our method outperforms the traditional ImageNet pretraining—both in supervised and unsupervised manner—by a clear margin, revealing the potential of semantic-based pretraining for further studies. In the future, we will focus on other downstream vision tasks with semantic-based VTBR, such as human parts segmentation and pose estimation.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable suggestions and constructive criticisms.

REFERENCES

[1] Zechen Bai, Zhigang Wang, Jian Wang, Di Hu, and Errui Ding. 2021. Unsupervised multi-source domain adaptation for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition. 12914–12923.

[2] Binghui Chen, Weihong Deng, and Jian Hu. 2019. Mixed high-order attention network for person re-identification. In IEEE International Conference on Computer Vision. 371–381.
[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In International Conference on Computer Vision and Pattern Recognition. IEEE, 248–255.

[4] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition. 994–1003.

[5] Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations. In Conference on Computer Vision and Pattern Recognition. 11162–11173.

[6] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. 2010. Person re-identification by symmetry-driven accumulation of local features. In IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2360–2367.

[7] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. 2021. Unsupervised pre-training for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition. 14750–14759.

[8] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S. Huang. 2019. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In International Conference on Computer Vision. 6112–6121.

[9] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. 2018. Fd-gan: Poseguided feature distilling gan for robust person re-identification. Advances in Neural Information Processing System 31 (2018).

[10] Kaiming He, Ross Girshick, and Piotr Dollár. 2019. Rethinking imagenet pre-training. In International Conference on Computer Vision. 4918–4927.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Conference on Computer Vision and Pattern Recognition. 770–778.

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017).

[13] Boseung Jeong, Jicheol Park, and Suha Kwak. 2021. Asmr: Learning attribute-based person search with adaptive semantic margin regularizer. In International Conference on Computer Vision. 12016–12025.

[14] Rithwik Kukunuri, Lucas Beyer, and Bastian Leibe. 2017. Lookahead optimizer: k steps forward, 1 step back. (2019).

[15] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In Conference on Computer Vision and Pattern Recognition. 152–159.

[16] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In Conference on Computer Vision and Pattern Recognition. 2197–2206.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European Conference on Computer Vision. Springer, 749–755.

[18] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. 2019. Improving person re-identification by attribute and identity learning. Pattern Recognition 95 (2019), 151–161.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In Conference on Computer Vision and Pattern Recognition. 3431–3440.

[20] Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Álvaro García-Martín. 2020. Semantic-aware scene recognition. Pattern Recognition 102 (2020), 107256.

[21] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In Conference on Computer Vision and Pattern Recognition Workshops. 0–0.

[22] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. 2021. Self-supervised pre-training for transformer-based person re-identification. arXiv preprint arXiv:2111.12084 (2021).

[23] Mohammed Bany Muhammad and Mohammed Yeasin. 2020. Eigen-CAM: Class activation map using principal components. In International Joint Conference on Neural Networks (IJCNN). IEEE, 1–7.

[24] Xin Ning, Ke Gong, Weijun Li, and Liping Zhang. 2021. JWSAA: Joint weak saliency and attention aware for person re-identification. Neurocomputing 453 (2021), 801–811.

[25] Xin Ning, Ke Gong, Weijun Li, Liping Zhang, Xiao Bai, and Shengwei Tian. 2020. Feature refinement and filter network for person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 31, 9 (2020), 3391–3402. DOI: 10.1109/TCSVT.2020.3043026

[26] Xingang Pan, Ping Luo, Jianping Shi, and Xiaou Tang. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In European Conference on Computer Vision. 464–479.

[27] Hyunjong Park and Bumsuk Ham. 2020. Relation network for person re-identification. In AAAI Conference on Artificial Intelligence. 11839–11847.
[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, SasaKoilis, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 32 (2019), 8026–8037.

[29] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaoqiang Gong, Tiejun Huang, and Yonghong Tian. 2016. Unsupervised cross-dataset transfer learning for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition. 1306–1315.

[30] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. 2018. Pose-normalized image generation for person re-identification. In European Conference on Computer Vision. 650–667.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning. PMLR, 8748–8763.

[32] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In European Conference on Computer Vision. Springer, 17–35.

[33] Suncheng Xiang, Yuzhuo Fu, Guanjie You, and Ting Liu. 2021. Taking a closer look at synthesis: Fine-grained attribute more: Learning from synthetic data with fine-grained attributes for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition. 6398–6407.

[34] Suncheng Xiang, Yuzhuo Fu, Guanjie You, and Ting Liu. 2020. Progression transfer learning for person re-identification. In IET Image Processing 14, 14 (2020), 3527–3535.

[35] Suncheng Xiang, Yuzhuo Fu, Mingye Xie, Zefang Yu, and Ting Liu. 2020. Unsupervised person re-identification by hierarchical cluster and domain transfer. Multimedia Tools and Applications 79, 27 (2020), 19769–19786.

[36] Suncheng Xiang, Yuzhuo Fu, Guanjie You, and Ting Liu. 2021. Taking a closer look at synthesis: Fine-grained attribute analysis for person re-identification. In IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 3765–3769.

[37] Suncheng Xiang, Dahong Qian, Mengyuan Guan, Binjie Yan, Ting Liu, Yuzhuo Fu, and Guanjie You. 2023. Less is more: Learning from synthetic data with fine-grained attributes for person re-identification. ACM Transactions on Multimedia Computing, Communications and Applications 19, 5s (2023), 1–20.

[38] Suncheng Xiang, Guanjie You, Lei Li, Mengyuan Guan, Ting Liu, Dahong Qian, and Yuzhuo Fu. 2022. Rethinking illumination for person re-identification: A unified view. In IEEE Conference on Computer Vision and Pattern Recognition. 4731–4739.

[39] QiQi Xiao, Hao Luo, and Chi Zhang. 2017. Margin sample mining loss: A deep learning based method for person re-identification. arXiv preprint arXiv:1710.00478 (2017).

[40] Zizheng Yang, Xin Jin, Kecheng Zheng, and Feng Zhao. 2022. Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition. 14298–14307.

[41] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. 2014. Deep metric learning for person re-identification. In International Conference on Pattern Recognition. IEEE, 34–39.
[51] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. 2020. Hierarchical clustering with hard-batch triplet loss for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition. 13657–13665.

[52] Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In Advances in Neural Information Processing Systems.

[53] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2014. Learning mid-level filters for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition. 144–151.

[54] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In IEEE International Conference on Computer Vision. 1116–1124.

[55] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J. Radke. 2019. Re-identification with consistent attentive siamese networks. In IEEE Conference on Computer Vision and Pattern Recognition. 5735–5744.

[56] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. 2019. Joint discriminative and generative learning for person re-identification. In IEEE Conference on Computer Vision and Pattern Recognition. 2138–2147.

[57] Zhidong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In IEEE International Conference on Computer Vision. 3754–3762.

[58] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. In IEEE Conference on Computer Vision and Pattern Recognition. 1318–1327.

Received 22 May 2023; revised 4 September 2023; accepted 7 October 2023