VTBR: SEMANTIC-BASED PRETRAINING FOR PERSON RE-IDENTIFICATION

Suncheng Xiang\textsuperscript{1}, Zirui Zhang\textsuperscript{1}, Mengyuan Guan\textsuperscript{1}, Hao Chen\textsuperscript{1}, Binjie Yan\textsuperscript{1}, Ting Liu\textsuperscript{1}, Yuzhuo Fu\textsuperscript{1}

\textsuperscript{1}School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China

ABSTRACT

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 ImageNet pretraining 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, we manually construct a diversified FineGPR-C caption dataset for the first time on person Re-ID events. Based on it, we propose a pure semantic-based pretraining approach named VTBR, which uses dense captions to learn visual representations with fewer images. Specifically, we train convolutional networks from scratch on the captions of FineGPR-C dataset, and transfer them to downstream Re-ID tasks. Comprehensive experiments conducted on benchmarks show that our VTBR can achieve competitive performance compared with ImageNet pretraining – despite using up to $1.4 \times$ fewer images, revealing its potential in Re-ID pretraining.

Index Terms— pretraining, re-identification, semantic-based, visual representation

1. INTRODUCTION

ImageNet pretraining 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 \cite{1} and segmentation \cite{2} on ImageNet \cite{3} dataset. In the field of person Re-ID, most of works \cite{4, 5, 6} 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 Fu et al. \cite{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 \cite{8, 9} 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 the academia and industry worldwide. For these reasons, there has been increasing interests for us to explore novel vision-and-language pretraining strategy which can replace the supervised ImageNet pretraining on Re-ID tasks. Unfortunately, existing datasets \cite{10, 11, 12, 13, 14, 15} in Re-ID community are all of limited scale due to the costly efforts required for data collection and annotation, especially none of them has diversified attributes to obtain dense captions, which cannot satisfy the need of semantic-based pretraining.

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 Fig. 1. Particularly, we jointly train a ResNet and transformers from scratch 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.

![Fig. 1. The overview of our 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.](image)

This work was supported by the National Natural Science Foundation of China under Project (Grant No.61977045). Preprint. Under review.
tions with better data-efficiency than other approaches. Another benefit of textual annotation is simplified data collection. Traditional labelling procedure of real pedestrian data always costs intensive human labor, sometimes even involving person privacy concerns and data security problems. In contrast, natural language description from fine-grained attributes on synthetic data do not require an explicit category and can be easily labelled by non-expert workers, leading to a simplified data labelling 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 the FineGPR-C caption dataset will serve as a solid baseline and advance the research of pretraining in Re-ID community.

As a consequence, the major contributions of our work can be summarized into three-fold: 1) We construct a FineGPR-C caption dataset for the first time to enable the semantic pretraining for Re-ID tasks. 2) Based on it, we propose a pure VirTex-based Re-ID pre-training approach named VTBR to learn visual representations from textual annotations. 3) Comprehensive experiments show that our VTBR matches or exceeds the performance of existing methods for supervised or unsupervised pretraining on ImageNet with fewer images.

2. PROPOSED METHOD

2.1. 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 dataset, which contains fine-grained attributes such as viewpoint, weather, illumination and background, as well as 13 accurate annotations at the identity level (shown in Fig. 2). 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 caption dataset. Some exemplars of FineGPR-C dataset are depicted in Fig. 3. It is worth mentioning that different pedestrian images have different captions by the different regular expressions.

During the caption generation based on fine-grained FineGPR dataset, we found that there exists serious redundancy among the different attributes. To maintain a high diversity of generated caption, we propose a Refined Selecting (RS) strategy to increase the inter-class diversity and minimize the intra-class variation of semantic caption. Particularly, we set a threshold and dynamically add the attribute which appears with a lower probability $\alpha$ into the final caption sentence $c$, the formula can be expressed as:

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

where $w_K, a_K$ denote fixed words and labelled attributes, respectively. $P_{a_K}$ represents the appearing probability of the attribute annotation in FineGPR. In summary, our goal is to improve the caption discriminative ability according to their attribute distribution, so the generated caption (token by token) will be more diversified and contain more discriminative

Fig. 2. (a) Illustration of attributes at the identity level; (b) Distribution of attributes in terms of IDs for FineGPR dataset.

Fig. 3. Some exemplars of semantic caption in FineGPR-C dataset, which is generated using our dynamic strategy.
2.2. Our 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 Fig. 4, our VTBR framework consists of a visual backbone ResNet-50 [17] and semantic backbone Transformer [18], 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 [8], 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 provides a learning signal to the visual backbone during pretraining. Note that we use the log-likelihood loss function to train the visual and semantic backbones in an end-to-end manner.

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

(2)

where \( \psi_f, \psi_b \) and \( \phi \) mean forward transformer, backward transformer and ResNet-50 respectively. \( T \) and \( V \) denote textual feature and visual feature separately. Therefore, we train our entire VTBR model from scratch without any pretrained weight 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 continuously. Specifically, we adopt global features extracted by visual backbone ResNet-50 to perform metric learning. Note that we only modify the output dimension of the latest fully-connected layer to the number of training identities [15]. During the period of testing, we extract the 2,048-dim pool-5 vector for retrieval under the Euclidean distance.

Fig. 4. The framework of our pretraining approach VTBR, which consists of a visual backbone ResNet-50 and Transformer. The visual backbone extracts visual features, and transformers extract semantic features to predict captions via bidirectional language modeling. After pretraining, the visual backbone is transferred to downstream Re-ID tasks.

### Table 1. Comparisons between ResNet and our VTBR pretraining on supervised tasks. “S” denotes semantic feature.

| Pretrain | Market-1501 | DukeMTMC | CUHK03 |
|----------|-------------|----------|--------|
| S #Imgs  | Rank-1 mAP | Rank-1 mAP | Rank-1 mAP |
| ResNet (ImageNet) | ×1.28M       | 94.3     | 86.7    | 61.6    |
| ResNet (FineGPR) | ×2.00M       | 85.5     | 59.8    | 37.8    |
| VTBR (FineGPR-C) | ✓1.83M       | 93.6     | 72.9    | 61.8    |
| VTBR+RS(FineGPR-C) | ✓0.91M       | 94.9     | 87.3    | 61.9    |

### 3. EXPERIMENTAL RESULTS

#### 3.1. Datasets
In this paper, we conduct experiments on three benchmarks, including Market-1501 [10], DukeMTMC-reID [11, 12] and CUHK03 [13]. Market-1501 has 1,501 identities in 32,668 images. 12,936 images of 751 identities are used for training, the query has 3,368 images and gallery has 19,732 images. DukeMTMC-reID contains 16,522 images of 702 identities for training, and the remaining images of 702 identities for testing. CUHK03 consists of 14,097 images with a total 1,467 identities. We evaluate the performance by mAP and Cumulative Matching Characteristic curves at Rank-1 and Rank-5.

#### 3.2. Implementation Details

For the pretraining of VTBR, we apply standard random cropping and normalization as data augmentation. Following the training procedure in [8], we adopt SGD with momentum 0.9 and weight decay \( 10^{-4} \) wrapped in LookAhead [19] with \( \alpha=0.5 \) and 5 steps. We empirically set the \( \alpha=0.8 \) in Eq. 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 follow a widely used open-source project\(^1\) as standard baseline, which is built only with commonly used softmax cross-entropy loss [20] and triplet loss.

\(^1\)https://github.com/michuanhaohao/reid-strong-baseline
loss [21] on vanilla ResNet-50. Following the practice in [22], the batch size of training samples is set as 64. As for triplet selection, we randomly selected 16 persons and sampled 4 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 [23] with two Nvidia GeForce RTX 3090 GPUs on a server equipped with a Intel Xeon Gold 6240 CPU.

3.3. Supervised Fine-tuning

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

3.4. Unsupervised Domain Adaption

Our semantic-based pretraining method enjoys the benefits of flexible corner scenarios of domain adaptive Re-ID tasks, where labelled 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 Table 2. When trained on Duke dataset, it can be easily observed that our VTBR+RS achieves a significant rank-1 performance of 50.6% and 5.7% on Market and CUHK03 respectively, outperforming the ImageNet pretraining by +2.6% and +0.8% in terms of rank-1 accuracy. When trained on Market dataset, our method can also lead to an obvious improvement of +1.9% on CUHK03 in rank-1 accuracy. However, when tested on Duke dataset, it is surprising to find that our method obtain a slightly inferior performance than ImageNet pretraining (mAP 13.5% vs. 13.8%, 0.91M vs. 1.28M images). We suspect that captions generated on FineGPR have obvious domain gap with Duke since there are some occlusion and multiple persons in the queries, which will undoubtedly degrade the performance of our method.

3.5. Visualization

In order to verify the effectiveness of our proposed method, we show some Grad-CAM [24] visualizations of some attention maps with VTBR in Fig. 5. 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 meaningful visual features with better semantic understanding, which significantly makes our semantic-based pretraining VTBR model more robust to perturbations.

4. CONCLUSION AND FUTURE WORK

This paper takes a big step forward to 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. In the future, we will focus on other downstream vision tasks with our VTBR, such as human related segmentation and pose estimation.
5. REFERENCES

[1] Kaiming He, Ross Girshick, and Piotr Dollár, “Rethinking imagenet pre-training,” in ICCV, 2019, pp. 4918–4927.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation,” in CVPR, 2015, pp. 3431–3440.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in CVPR. Ieee, 2009, pp. 248–255.

[4] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang, “Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification,” in ICCV, 2019, pp. 6112–6121.

[5] Zechen Bai, Zhigang Wang, Jian Wang, Di Hu, and Errui Ding, “Unsupervised multi-source domain adaptation for person re-identification,” in CVPR, 2021, pp. 12946–12954.

[6] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo, “Hierarchical clustering with hard-batch triplet loss for person re-identification,” in CVPR, 2020, pp. 13657–13665.

[7] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen, “Unsupervised pre-training for person re-identification,” in CVPR, 2021, pp. 14730–14739.

[8] Kaiming He, Ross Girshick, and Piotr Dollár, “Rethinking imagenet pre-training,” in ICCV, 2019, pp. 4918–4927.

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in CVPR, 2014, pp. 152–159.

[10] Suncheng Xiang, Yuzhuo Fu, Guanshuo Wang, and Wei Jiang, “Taking a closer look at synthesis: Fine-grained attribute analysis for person re-identification,” in ICASSP. IEEE, 2021, pp. 3765–3769.

[11] Suncheng Xiang, Guanshuo Wang, and Wei Jiang, “Bag of tricks and a strong baseline for deep person re-identification,” in CVPRW, 2019, pp. 0–0.

[12] Zheng Zhang, Liang Zheng, and Yi Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in ICCV, 2017, pp. 3754–3762.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in NIPS, 2017, pp. 5998–6008.

[14] Rithwik Kukunuri and Shivji Bhagat, “Lookahead optimizer: k steps forward, 1 step back,” 2019.

[15] Zhilu Zhang and Mert R Sabuncu, “Generalized cross entropy loss for training deep neural networks with noisy labels,” in NIPS, 2018.

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

[17] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang, “Bag of tricks and a strong baseline for deep person re-identification,” in CVPRW, 2019, pp. 0–0.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., “Pytorch: An imperative style, high-performance deep learning library,” NIPS, vol. 32, pp. 8026–8037, 2019.

[19] Mohammed Bany Muhammad and Mohammed Yeasin, “Eigen-cam: Class activation map using principal components,” in 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020, pp. 1–7.