PVTNL: Prompting Vision Transformers with Natural Language for Generalizable Person Re-identification

Ning Wang[1] **Lei Xie**[2] * **Shiwei Gan**[3] **Sanglu Lu**[4]
State Key Laboratory for Novel Software Technology, Nanjing University, China,
{ning_wang[1], sw[3]}@smail.nju.edu.cn {lxie[2], sanglu[4]}@nju.edu.cn

Abstract

Domain generalization person re-identification (DG ReID) aims to train models on source domains and generalize to unseen target domains. While patch-based Vision Transformers have achieved success in capturing fine-grained visual features, they often overlook global semantic structure and suffer from feature entanglement, leading to overfitting across domains. Meanwhile, natural language provides high-level semantic abstraction but lacks spatial precision for fine-grained alignment. We propose PVTNL (Prompting Vision Transformers with Natural Language), a novel framework for generalizable person re-identification. PVTNL leverages the pre-trained vision-language model BLIP to extract aligned visual and textual embeddings. Specifically, we utilize body-part cues to segment images into semantically coherent regions and align them with corresponding natural language descriptions. These region-level textual prompts are encoded and injected as soft prompts into the Vision Transformer to guide localized feature learning. Notably, our language module is retained during inference, enabling persistent semantic grounding that enhances cross-domain generalization. Extensive experiments on standard DG ReID benchmarks demonstrate that PVTNL achieves state-of-the-art performance. Ablation studies further confirm the effectiveness of body-part-level alignment, soft language prompting, and the benefit of preserving language guidance at inference.

1 Introduction

Person Re-Identification (ReID) aims to retrieve individuals with the same identity across a database, where the images are captured under different cameras, timestamps, and spatial locations (Zheng et al., 2015, 2017; Chen et al., 2019a; Zhu et al., 2020; Dou et al., 2022). As a crucial task in computer vision, ReID has been widely applied in video surveillance (Wang et al., 2021) and security systems (Jiang and Ye, 2023). With the rapid development of convolutional neural networks (CNNs) (Krizhevsky et al., 2012; Tan and Le Q V, 1905), ReID has achieved impressive progress (Li et al., 2023a; Lian et al., 2023; Luo et al., 2019). However, when the deployment scenario differs significantly from the training domain, performance degrades dramatically due to domain overfitting (Carlucci et al., 2019). To tackle this challenge, Domain Generalization Person Re-Identification (DG ReID) has been proposed to assess the model's robustness under domain shifts without accessing target domain data during training (Yi et al., 2014; Jia et al., 2019; Li et al., 2020; Dai et al., 2021; Gu et al., 2023).

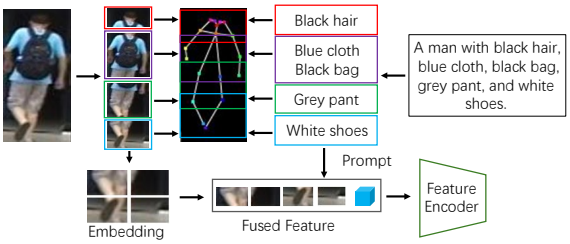

Figure 1: **Demonstration of our model.** We segment the image into body-part-based local regions and align each region with its corresponding natural language description. The textual descriptions are encoded and injected as prompts to assist local feature extraction within each region.

Existing DG ReID methods have explored various directions, including domain-invariant feature learning (Yuan et al., 2020; Zhang et al., 2022c), domain disentanglement (Jin et al., 2020), domain alignment (Choi et al., 2021; Han et al., 2023; Jiao et al., 2022), and meta-learning (Ni et al., 2022; Zhao et al., 2021; Zhang et al., 2023). While these techniques have shown promising results, they often rely on distribution similarity between domains and may yield unstable generalization performance

*Lei Xie is the corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21663–21674
November 4-9, 2025

(Robinson et al., 2021; Nichol et al., 2018; Rangwani et al., 2022).

Meanwhile, Vision Transformers (ViTs) (Vaswani et al., 2017) have demonstrated stronger robustness to distribution shifts compared to CNNs. Several works have leveraged ViTs in DG ReID (Zhang et al., 2022a; He et al., 2021; Li and Gong, 2025; Cho et al., 2024), achieving enhanced generalization. However, pure visual models often struggle to distinguish fine-grained semantics under complex conditions such as varying poses or accessories.

To address this limitation, we propose PVTNL (Prompting Vision Transformers with Natural Language), a novel framework that introduces natural language prompts into ViTs for region-aware feature learning, as shown in Fig. 1. Specifically, we first use body-part cues to segment each image into semantically consistent regions and align them with corresponding textual descriptions. These region-level texts are encoded and injected into the Vision Transformer as soft prompts, guiding the network to attend to meaningful local features. Additionally, we employ a cross-attention mechanism to fuse local features, further enhancing the model's representational capacity. Unlike prior works, our method retains the natural language module during inference, providing persistent semantic grounding that improves cross-domain generalization. PVTNL achieves state-of-the-art performance on several DG ReID benchmarks, offering a new perspective on incorporating language as a generalization prior in vision tasks.

The key contributions of this paper are summarized as follows:

a) We propose PVTNL, a novel framework that integrates natural language prompts into Vision Transformers, achieving persistent semantic grounding and improved domain generalization.

b) We introduce a region-level alignment mechanism based on body-part cues and natural language descriptions, enhancing the model's ability to capture high-level semantic features.

c) Extensive experiments on multiple benchmark datasets demonstrate that PVTNL achieves state-of-the-art performance on DG ReID tasks.

2 Related Work

2.1 Person Re-identification

Person Re-Identification (ReID) aims to retrieve individuals across camera views from non-overlapping locations (Zheng et al., 2016; Sun et al., 2018; Chen et al., 2019b; Ye et al., 2021; Li et al., 2021). Early approaches (Farenzena et al., 2010; Ding et al., 2015) focused on designing handcrafted or learned features to match pedestrian images. With the rise of deep learning frameworks such as PyTorch (Paszke et al., 2019), convolutional neural networks (CNNs) have become the dominant paradigm for ReID, owing to their powerful feature extraction capabilities. For example, (Xiang et al., 2020) proposed a metric-learning-based framework using CNNs to extract robust features, while (Chen et al., 2017) introduced a local feature approach with enhanced receptive fields via multi-scale downsampling. Additionally, (Gu et al., 2022) presented a clothing-invariant adversarial loss to extract identity-consistent features from RGB images. Although these CNN-based methods achieve impressive performance on common benchmarks, their generalization capability degrades significantly when exposed to domain shifts, such as changes in environment, camera style, or clothing.

2.2 Domain Generalization Person ReID

Domain Generalization Person ReID (DGReID) (Zheng et al., 2016; Song et al., 2019; Ye et al., 2021; Qi et al., 2022; Xie et al., 2024) focuses on improving a model's ability to generalize to unseen target domains without accessing target data during training. Given that real-world deployment scenarios are diverse and not covered by any single dataset, DGReID holds high practical value. First introduced by (Yi et al., 2014), subsequent works such as (Choi et al., 2021; Song et al., 2019) have explored learning domain-invariant representations through adversarial training. Others, like (Xu et al., 2022), proposed normalization-based alignment strategies to mitigate domain gaps, while (Zhao et al., 2021) employed meta-learning to simulate domain shifts during training. Despite their progress, most of these methods rely heavily on discriminative or contrastive learning (Dou et al., 2023; Jin et al., 2020), making them sensitive to the distribution similarity between source and target domains. When a significant domain gap exists, these approaches tend to fail to generalize effectively, thus limiting further improvements.

2.3 Vision Transformer for Person ReID

The Vision Transformer (ViT) (Dosovitskiy et al., 2020) introduced the Transformer architecture (Vaswani et al., 2017) into computer vision by

replacing convolutions with self-attention mechanisms. ViTs naturally model long-range dependencies and exhibit stronger generalization capabilities under distributional shifts. (He et al., 2021) was the first to adapt ViT to the ReID task. Following this, (Liao and Shao, 2021) replaced standard self-attention with cross-attention to better capture pairwise feature similarities. However, despite these advances, most ViT-based methods rely on rigid patch-based partitioning and focus predominantly on low-level visual patterns. As a result, they struggle to handle perspective variations and fine-grained semantic ambiguities, such as subtle differences in clothing or pose (Ni et al., 2023).

2.4 Vision-Language Models and Prompt-based Methods

Vision-Language Models (VLMs) aim to jointly process visual and textual information, and have shown remarkable potential in tasks such as image captioning, cross-modal retrieval, and multimodal reasoning. Recent works such as LLaVA (Li et al.) and Qwen-VL (Bai et al., 2023) focus on real-world visual understanding, while models like CLIP (Radford et al., 2021), BLIP (Li et al., 2022, 2023b), and MiniGPT (Zhu et al., 2023; Chen et al., 2023) learn strong cross-modal alignment by contrastive or generative training.

Prompt-based methods have also been introduced into person ReID. For instance, CLIP-ReID (Li et al., 2023c) leverages CLIP to align textual and visual features, and ReFID (Peng et al., 2024) applies prompt tuning to adapt VLMs for cross-domain ReID. Other recent approaches such as PAT (Ni et al., 2023) and Prompt-CLIP (Cui et al., 2025) explore domain generalization by injecting textual prompts. While effective, these methods typically operate at the global image level and overlook fine-grained body-part semantics.

Our work differs from these prior approaches in three key aspects. First, unlike generic prompt strategies, we leverage structured human pose priors to construct body-part-level prompts that provide fine-grained semantic grounding. Second, in contrast to existing vision–language alignment and segmentation methods, we explicitly couple pose-guided segmentation with language prompts to improve domain robustness. Finally, unlike prompt-based models in other domains such as scene graph generation (SGG) (Li et al., 2024), which focus on logical or relational alignment, our method is specifically optimized for DG-ReID, targeting improved interpretability and performance under domain shifts.

3 Methodology

In this section, we introduce **PVTNL** (Prompting Vision Transformers with Natural Language), a novel framework designed for person re-identification by integrating local vision prompts with structured natural language. As shown in Fig. 2, PVTNL consists of four main modules: i) image segmentation and language-guided alignment; ii) prompt injection; iii) local features fusion; and iv) the overall training loss design. We describe each component in detail below.

3.1 Image Segmentation and Language-Guided Alignment

Image Segmentation. Since our model focuses on region-level features, to enable fine-grained reasoning at the part level, we first segment each person image into semantically consistent body regions. Given an input image $x \in \mathbb{R}^{H\times W\times C}$ and its 18 annotated pose landmarks $P = \{(x_1, y_1), (x_2, y_2), ..., (x_{18}, y_{18})\} \in \mathbb{R}^{18\times 2}$, we define four body part regions PR_i, $i \in \{1, 2, 3, 4\}$: head, upper torso, lower body, and feet. The body keypoints are obtained using the HRNet pose estimation model trained on the COCO keypoint dataset (Sun et al., 2019).

- **Head Region** (PR_1): Includes nose, neck, eyes, ears, and shoulders.
- **Upper Torso** (PR_2): Includes shoulders, elbows, wrists, and hips.
- **Lower Body** (PR_3): Includes hips, knees, and ankles.
- **Feet Region** (PR_4): Includes ankles and feet.

To transform the discrete keypoints into continuous bounding boxes, we compute the maximum and minimum coordinates of the keypoints within each region and expand them by a margin c:

$$PR_i = [x_{\min} - c, x_{\max} + c, y_{\min} - c, y_{\max} + c] \quad (1)$$

The cropped sub-image corresponding to region PR_i is denoted as x_{PR_i}.

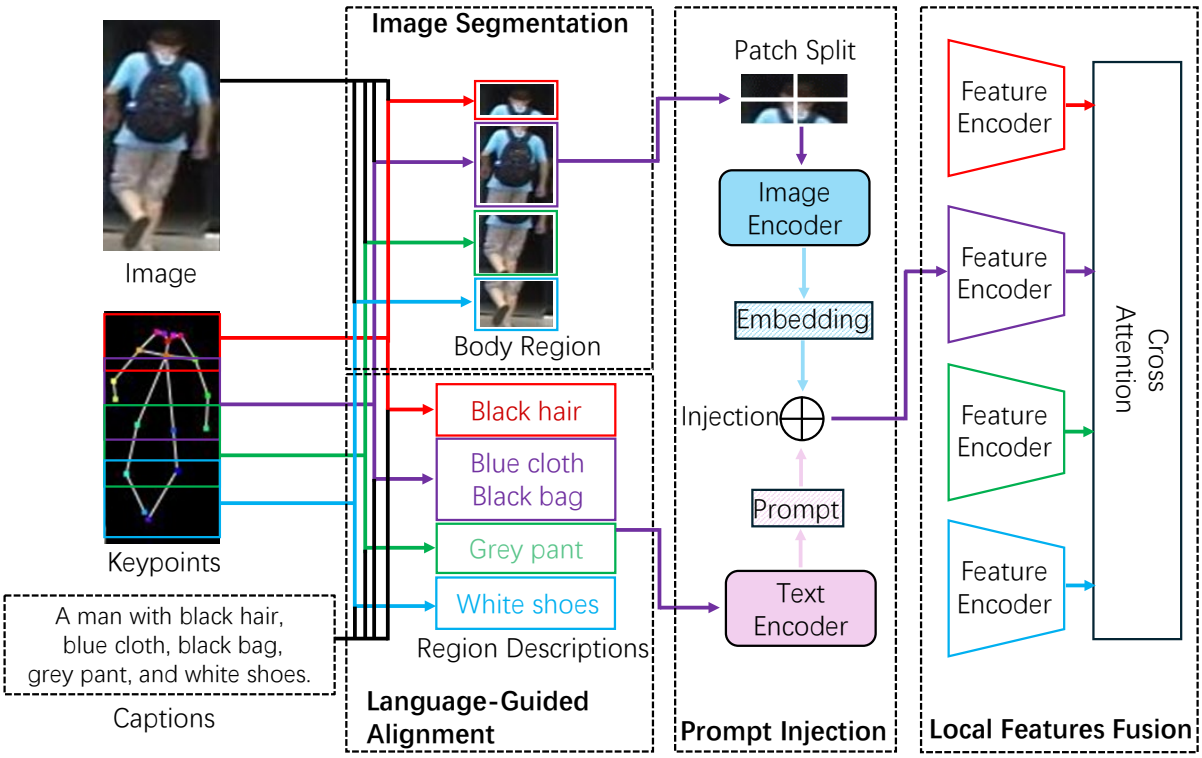

Figure 2: **Overview of PVTNL.** We leverage body-part keypoints to segment the input image into semantically consistent regions. Each region is aligned with corresponding natural language descriptions extracted from the image. These textual descriptions are then encoded as prompts and injected into the Vision Transformer to guide local region-specific feature learning.

Language-Guided Alignment. We adopt a BLIP-based vision-language pre-trained model to generate image captions. Since cropped regions lose contextual semantics, we generate a full image-level caption D using the original image x. Then, we extract region-specific descriptions RD_i by matching pre-defined keyword sets RDK_i:

$$
\begin{aligned}
RDK_1 &= \{\text{Hair, Hat}\} \\
RDK_2 &= \{\text{Cloth, Top, Vest, Bag}\} \\
RDK_3 &= \{\text{Pants, Skirt}\} \\
RDK_4 &= \{\text{Shoes, Boots}\}
\end{aligned}
$$

Based on these keywords, we parse D into structured region-level descriptions RD_i, where $i \in \{1, 2, 3, 4\}$.

3.2 Prompt Injection

Patch Embedding of Part Regions. Each region image x_{PR_i} is divided into M non-overlapping patches, and embedded as:

$$\mathcal{E}_i = [x_{\text{cls}_i}, x_{i_1}, x_{i_2}, ..., x_{i_M}] + \mathcal{P} \tag{2}$$

where x_{cls_i} is a learnable class token and $\mathcal{P}$ denotes position embeddings.

Textual Prompt Extraction. We use the BLIP text encoder to extract region prompt embeddings from RD_i. The textual prompt is projected to the same dimension as the visual patch embeddings:

$$\text{Prompt}_i = \text{TextEncoder}(RD_i) \tag{3}$$

Prompt Injection and Feature Extraction. We inject the textual prompt into the part image representation via early fusion. Specifically, the fused sequence is:

$$\mathcal{Z}_i = [\mathcal{E}_i, \text{Prompt}_i] \tag{4}$$

We compute attention parameters from $\mathcal{Z}_i$:

$$Q_i = \mathcal{Z}_i W_Q = [q_{\text{cls}_i}, q_{i_1}, ..., q_{i_M}, q_{p_i}] \tag{5}$$

$$K_i = \mathcal{Z}_i W_K = [k_{\text{cls}_i}, k_{i_1}, ..., k_{i_M}, k_{p_i}] \tag{6}$$

$$V_i = \mathcal{Z}_i W_V = [v_{\text{cls}_i}, v_{i_1}, ..., v_{i_M}, v_{p_i}] \tag{7}$$

where W_Q, W_K, and W_V are learnable linear projections. The attention is computed as:

$$\text{Attention}(Q_i, K_i, V_i) = \text{Softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}}\right) V_i \tag{8}$$

The resulting fused local feature is $f(PR_i)$.

Computational Cost. Our method introduces negligible additional computational overhead. For an input image of size 256×128 with a patch size of 16, the ViT-B backbone produces 128 visual tokens. We concatenate 20 textual tokens (each of dimension 768) as prompts. This results in an additional $\sim$51M FLOPs compared to the baseline ViT-B (17.6G FLOPs), corresponding to only $\sim$0.29% overhead, which is practically insignificant.

3.3 Contrastive Learning Loss

To encourage local part representations to be discriminative, we adopt a contrastive loss. Let $f(x_{ij})$ be the i-th part region feature of the j-th image. We retrieve k nearest positive samples D_{ij}^{+} from a memory dictionary, and define the loss as:

$$\mathcal{L}_{ij}^{\text{contrast}} = -\log \frac{\exp\left(f(x_{ij})^{\top} D_{ij}^{+}/\tau\right)}{\sum_{d \in D_{ij}} \exp\left(f(x_{ij})^{\top} d/\tau\right)} \tag{9}$$

where τ is a temperature hyperparameter.

3.4 Fusion of Local Features

After obtaining the region-level features, we fuse them to derive a global representation using cross-attention across regions. For part regions i and j, we compute:

$$\text{CrossAttention}(Q_i, K_j, V_j) = \text{Softmax}\left(\frac{Q_i K_j^{\top}}{\sqrt{d_k}}\right) V_j \tag{10}$$

By aggregating all pairwise cross-attention, we obtain the global feature $g(x_j)$ for the j-th image.

3.5 Overall Training Loss Function

In addition to the contrastive loss, we use the triplet loss to enforce ranking constraints between positive and negative samples:

$$\mathcal{L}_{\text{tri}} = \max\left(0, d_{a,p} - d_{a,n} + m\right) \tag{11}$$

where $d_{a,p}$ and $d_{a,n}$ denote the distances between anchor-positive and anchor-negative, respectively, and m is a margin hyperparameter.

The total loss is a weighted combination:

$$\mathcal{L}_{\text{total}} = \lambda_1 \sum_{i,j} \mathcal{L}_{ij}^{\text{contrast}} + \lambda_2 \mathcal{L}_{\text{tri}} \tag{12}$$

Table 1: Statistics of public Re-ID datasets.

Dataset	#ID	#Image	#Camera
Market-1501 (Zheng et al., 2015)	1,501	32,217	6
MSMT17 (Wei et al., 2018)	4,101	126,441	15
CUHK02 (Li and Wang, 2013)	1,816	7,264	10
CUHK03 (Li et al., 2014)	1,467	14,096	2
CUHK-SYSU (Xiao et al., 2017)	11,934	34,574	1
PRID (Hirzer et al., 2011)	200	1,134	2
GRID (Loy et al., 2009)	250	1,275	8
VIPeR (Gray and Tao, 2008)	632	1,264	2
iLIDs (Wang et al., 2014)	119	476	2

4 Experiments

4.1 Datasets and Evaluation Metrics

We conducted experiments on the following datasets: Market-1501 (Zheng et al., 2015), MSMT17 (Wei et al., 2018), CUHK02 (Li and Wang, 2013), CUHK03 (Li et al., 2014), CUHK-SYSU (Xiao et al., 2017), PRID (Hirzer et al., 2011), GRID (Loy et al., 2009), VIPeR (Gray and Tao, 2008), and iLIDs (Wang et al., 2014). The statistics of the datasets are shown in Table 1. To simplify the notation, we use M to represent Market-1501, MS to represent MSMT17, C2 to represent CUHK02, C3 to represent CUHK03 and CS to represent CUHK-SYSU.

To evaluate domain generalization (DG) ReID, we follow two common protocols:

- **Protocol 1:** Train on M, C2, C3, and CS; test on PRID, GRID, VIPeR, and iLIDs.
- **Protocol 2:** Train on M, MS, C3, and CS, using a leave-one-out strategy where one dataset is used for testing and the others for training.

The evaluation metrics used are the widely recognized Cumulative Matching Characteristics (CMC) at Rank-1 and the mean average precision (mAP).

4.2 Implementation Details

We implemented our model in PyTorch (Paszke et al., 1912) and conducted training on an RTX-3090 GPU. For the backbone, we utilized the pre-trained ViT-B16 (Dosovitskiy et al., 2020) with a patch size of 16. The batch size for training samples was set to 64, and the input images were resized to 256×128. For vision-language processing, we utilize a BLIP-based vision-language pre-trained model (Li et al., 2022) for both image caption generation and text encoding. We applied standard data augmentation techniques, including random flipping, cropping, and color jittering (Gong, 2021).

Table 2: Comparison with state-of-the-art. We use **bold** to indicate the best result and underline to indicate the second-best result. The methods marked by "*" are unsupervised.

Method	Reference	Protocol 1									
		PRID		GRID		VIPeR		iLIDS		Average	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
MoCo*	CVPR2020	10.9	6.5	6.9	2.8	7.5	4.0	46.4	38.8	17.9	13.0
LUP*	CVPR2021	3.7	1.5	4.0	1.2	5.0	1.4	43.0	36.7	13.9	10.2
LUPnl*	CVPR2022	12.2	8.1	7.4	3.1	9.2	4.6	49.8	43.3	19.7	14.8
CrossGrad	arXiv2018	28.2	18.8	16.0	9.0	30.4	20.9	61.3	49.7	34.0	24.6
MLDG	AAAI2018	35.4	24.0	23.6	15.8	33.5	23.5	65.2	53.8	39.4	29.3
PPA	CVPR2018	45.3	31.9	38.0	26.9	54.5	45.1	72.7	64.5	52.6	42.1
DIMN	CVPR2019	52.0	39.2	41.1	29.3	60.1	51.2	78.4	70.2	57.9	47.5
SNR	CVPR2020	66.5	52.1	47.7	40.2	61.3	52.9	89.9	84.1	66.4	57.3
QAConv	ECCV2020	62.2	52.3	57.4	48.6	66.3	57.0	81.9	75.0	67.0	58.2
DML	CVPR2021	60.4	47.3	49.0	39.4	58.0	49.2	84.0	77.3	62.9	53.3
M^3L	CVPR2021	65.3	55.0	50.5	40.0	68.2	60.8	74.3	65.0	64.6	55.2
ViT-B	ICLR2021	63.8	52.0	56.0	44.8	74.8	65.8	76.2	65.0	67.7	56.9
Trans	ICCV2021	68.1	59.0	60.8	49.6	69.5	60.1	79.8	68.3	69.6	59.3
MetaBIN	CVPR2021	70.8	61.2	57.9	50.2	64.3	55.9	82.7	74.7	68.9	60.5
DDAN	AAAI2021	67.5	62.9	50.9	46.2	60.8	56.5	81.2	78.0	65.1	60.9
RaMoE	CVPR2021	67.3	57.7	54.2	46.8	64.6	56.6	90.2	85.0	69.1	61.5
MDA	CVPR2022	-	-	62.9	61.2	71.7	63.5	84.4	80.4	-	-
META	ECCV2022	71.7	61.9	60.1	52.4	68.4	61.5	83.5	79.2	70.9	63.8
ACL	ECCV2022	73.4	63.0	65.7	55.2	75.1	66.4	86.5	81.8	75.2	66.6
PAT	ICCV2023	57.9	46.0	54.5	45.6	67.8	60.1	78.1	66.7	64.6	54.6
CLIP	AAAI2023	68.3	57.0	58.2	48.8	69.3	60.1	83.4	75.0	69.8	60.2
ReFID	TOMM2024	71.3	63.2	59.8	56.1	68.7	60.9	84.6	81.0	71.1	65.3
GMN	TCSVT2024	75.4	66.0	64.8	54.4	77.7	69.0	-	-	-	-
Ours(PVTNL)	This Paper	**78.7**	**71.2**	**75.3**	**67.0**	**82.2**	**74.7**	**91.3**	**89.6**	**81.9**	**75.6**

The optimizer used for the model was SGD with an initial learning rate of 1×10^{-3}, which decayed gradually during the training process. The temperature hyperparameter τ is set to 0.3, and the region margin c is defined as 10% of the region's width and height. We adopt a two-stage training strategy: first, we freeze the BLIP encoder and optimize the image encoder using contrastive loss; then, we freeze both encoders and fine-tune the fusion layers using triplet loss.

4.3 Comparison with State-of-the-Arts

To validate our method, we compare our model with state-of-the-art (SOTA) methods, including MoCo (He et al., 2020), LUP (Fu et al., 2021), LUPnl (Fu et al., 2022), CrossGrad (Shankar et al., 2018), MLDG (Li et al., 2018), PPA (Qiao et al., 2018), DIMN (Song et al., 2019), SNR (Jin et al., 2020), QAConv (Liao and Shao, 2020), DML (Dai et al., 2021), M^3L (Zhao et al., 2021), ViT-B (Dosovitskiy et al., 2020), Trans (He et al., 2021), MetaBIN (Choi et al., 2021), DDAN (Chen et al., 2021), RaMoE (Dai et al., 2021), MDA (Ni et al., 2022), META (Xu et al., 2022), ACL (Zhang et al., 2022b), PAT (Ni et al., 2023), CLIP (Li et al., 2023c), ReFID (Peng et al., 2024), GMN (Qi et al., 2024), and CycAs (Wang et al., 2020). As shown in Table 2 and Table 3, our model achieves the best results under both protocols. This demonstrates that our model effectively accomplishes the DG ReID task.

Under protocol 1, our method achieve the best performance, as shown in Table 2. In particular, on the PRID dataset, we achieved mAP of 78.7% and R1 of 71.2%, surpassing the performance of GMN by 3.3% on mAP and 5.2% on R1. On the GRID, VIPeR and iLIDS datasets, our method outperforms these models with mAP of 75.3%, 82.2% and 91.3%. The average performance on all four datasets of our method is 81.9% on mAP and 75.6% on R1, surpassing the performance of ACL by 6.7% on mAP and 9.0% on R1.

Under protocol 2, we also conducted extensive experiments to validate our model. The results are

Table 3: Comparison with state-of-the-art. We use **bold** to indicate the best result and underline to indicate the second-best result. The methods marked by "*" are unsupervised.

Method	Reference	Protocol 2							
		Market1501		MSMT17		CUHK03		Average	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1
LUP*	CVPR2021	1.0	3.3	0.1	0.3	0.5	0.1	0.5	1.2
MoCo*	CVPR2020	2.6	10.5	0.2	0.5	0.7	0.3	1.2	3.8
LUPnl*	CVPR2022	3.8	13.8	0.2	0.6	0.8	0.4	1.6	4.9
CycAs*	arXiv2022	57.5	80.3	20.2	43.9	26.5	25.8	34.7	50
QAConv	ECCV2020	63.1	83.7	16.4	45.3	25.4	24.8	35.0	51.3
DML	CVPR2021	49.9	75.4	9.9	24.5	32.6	32.9	30.8	44.3
MetaBIN	CVPR2021	57.9	80.1	17.8	40.2	28.8	28.1	34.8	49.5
RaMoE	CVPR2021	56.5	82.0	13.5	34.1	35.5	36.6	35.2	50.9
M^3L	CVPR2021	61.5	82.3	16.7	37.5	34.2	34.4	37.5	51.4
ViT-B	ICLR2021	59.2	78.3	20.5	42.7	36.5	35.8	38.7	52.3
Trans	ICCV2021	59.9	79.8	23.2	46.3	36.5	36.1	39.9	54.1
META	ECCV2022	67.5	86.1	22.5	49.9	36.3	35.1	42.1	57.0
ACL	ECCV2022	74.3	89.3	20.4	45.9	41.2	41.8	45.3	59.0
CLIP	AAAI2023	68.8	84.4	26.6	53.1	42.1	41.9	45.8	59.8
ReFID	TOMM2024	67.6	85.3	18.3	39.8	33.3	34.8	39.7	53.3
GMN	TCSVT2024	72.3	87.1	24.4	50.9	43.2	42.1	46.6	60.0
Ours(PVTNL)	This Paper	**76.8**	**90.4**	**30.3**	**58.1**	**47.2**	**46.5**	**51.4**	**65.0**

Table 4: Ablation studies of our method. Where the "V" represents ViT-B; "S" represents Image Segmentation; "P" represents Prompt Injection; "F" represents Feature Fusion.

Method	Protocol 1		Protocol 2	
	mAP	R1	mAP	R1
V (w/o S,P,F)	67.7	56.9	38.7	52.3
V+S (w/o P,F)	73.4	62.8	45.6	58.1
V+S+P (w/o F)	78.2	69.8	48.4	63.4
V+S+P+F	81.9	75.6	51.4	65.0

shown in Table 3. Although some latest works, such as GMN and CLIP, have achieved good performances, where the average performances are 46.6% on mAP and 60.0% on R1 for GMN and 45.8% on mAP and 59.8% on R1 for CLIP, our method surpasses them and achieves the best results. The average performance on all three datasets of our method is 51.4% on mAp and 65.0% on R1, surpassing the performance of GMN by 4.8% on mAP and 5.0% on R1.

4.4 Ablation Studies

Our model consists of three key components: Image Segmentation, Prompt Injection, and Feature Fusion. To evaluate their individual contributions, we conduct ablation studies under both Protocol 1 and Protocol 2. Specifically, we consider the following settings:

1. Replace the Image Segmentation module with a vanilla ViT-B model using full images.

2. Remove the Prompt Injection component, retaining only segmentation and part region extraction.

3. Replace the Feature Fusion module with simple feature concatenation.

As shown in Table 4, each component contributes significantly to the model's performance. The segmentation module effectively partitions regions with consistent semantics, prompt injection provides textual guidance, and feature fusion integrates fine-grained local features into a global representation. When combined, our full PVTNL model achieves 81.9% mAP and 75.6% Rank-1 in Protocol 1, and 51.4% mAP and 65.0% Rank-1 in Protocol 2.

Table 5: Hyperparameter experiments for part regions.

Number of part regions	Protocol 1		Protocol 2	
	mAP	R1	mAP	R1
2 (head & torso, legs & feet)	77.6	68.9	47.5	61.2
3 (head, torso, lower body)	80.6	73.4	48.7	63.3
4 (head, upper, lower, feet)	81.9	75.6	51.4	65.0

4.5 Hyperparameter Experiments

We further explore how the number of segmented regions affects performance. As shown in Table 5, increasing the number of part regions from 2 (e.g., head & torso, legs & feet), to 3 (head, torso, lower body), and finally to 4 (head, upper, lower, feet), improves the model's accuracy. This is because finer segmentation allows the model to extract more discriminative features, enhancing its ability to distinguish between identities.

5 Conclusion

In this paper, we propose **PVTNL** (Prompting Vision Transformers with Natural Language), a novel framework that integrates natural language prompts into Vision Transformers for region-aware feature learning in domain generalizable person re-identification (DG ReID). We use body-part cues to segment each image into semantically consistent regions and align them with corresponding textual descriptions. We encode textual descriptions into prompts, inject them into the Vision Transformer for local feature extraction, and employ cross-attention to fuse local features for enhanced representation learning. Experimental results across multiple datasets and different evaluation protocols demonstrate that our model achieves state-of-the-art performance in DG-ReID.

In future work, we will extend PVTNL to cross-modal scenarios (e.g., visible–infrared person ReID) and explore its deployment under ethical guidelines to ensure responsible use.

6 Limitations

Our model relies on body-part keypoints to align local image regions with textual descriptions. When the keypoints are inaccurately detected, it can cause misalignment between regions and text features, degrading performance. To mitigate this, we apply a fallback mechanism that excludes regions with low-confidence keypoints, and we adopt a spatial expansion strategy to improve robustness against moderate pose estimation noise.

In addition, because we utilize BLIP as the image captioning model, the generated textual descriptions may sometimes exceed the semantic scope of the predefined region-specific keywords. This discrepancy can make it difficult to retrieve the correct regional textual features, further affecting alignment quality.

Although our benchmarks already span diverse domains (e.g., indoor vs. outdoor, varying lighting and camera styles), extending our approach to cross-modal scenarios such as visible–infrared person ReID remains a valuable future direction.

Finally, person ReID is a high-risk application that raises potential ethical concerns, such as privacy invasion, surveillance misuse, and bias amplification. We emphasize that our method is intended for research purposes, and its deployment should be carefully governed by ethical guidelines and legal regulations to ensure responsible use.

7 Acknowledgements

This work is supported in part by National Natural Science Foundation of China under Grant No. 92467202; Natural Science Foundation of Jiangsu Province (Key Program) under Grant No. BK20243040; National Natural Science Foundation of China under Grant Nos. 62272216, 62372224; and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

J Bai, S Bai, S Yang, S Wang, S Tan, P Wang, J Lin, C Zhou, and J Qwen-VL Zhou. 2023. A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2229–2238.

Binghui Chen, Weihong Deng, and Jiani Hu. 2019a. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 371–381.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Peixian Chen, Pingyang Dai, Jianzhuang Liu, Feng Zheng, Mingliang Xu, Qi Tian, and Rongrong Ji. 2021. Dual distribution alignment network for generalizable person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1054–1062.

Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. 2019b. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8351–8361.

Yanbei Chen, Xiatian Zhu, and Shaogang Gong. 2017. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2590–2600.

Yoonki Cho, Jaeyoon Kim, Woo Jae Kim, Junsik Jung, and Sung-eui Yoon. 2024. Generalizable person re-identification via balancing alignment and uniformity. *arXiv preprint arXiv:2411.11471*.

Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. 2021. Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3425–3435.

Fangming Cui, Yonggang Zhang, Xuan Wang, Xule Wang, and Liang Xiao. 2025. Generalizable prompt learning of clip: A brief overview. *arXiv preprint arXiv:2503.01263*.

Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. 2021. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16145–16154.

Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Zhaopeng Dou, Zhongdao Wang, Weihua Chen, Yali Li, and Shengjin Wang. 2022. Reliability-aware prediction via uncertainty learning for person image retrieval. In *European Conference on Computer Vision*, pages 588–605. Springer.

Zhaopeng Dou, Zhongdao Wang, Yali Li, and Shengjin Wang. 2023. Identity-seeking self-supervised representation learning for generalizable person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15847–15858.

Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. 2010. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2360–2367. IEEE.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14750–14759.

Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. 2022. Large-scale pre-training for person re-identification with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2476–2486.

Yunpeng Gong. 2021. A general multi-modal data learning method for person re-identification. *arXiv preprint arXiv:2101.08533*, 1(3):4.

Douglas Gray and Hai Tao. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, pages 262–275. Springer.

Jianyang Gu, Hao Luo, Kai Wang, Wei Jiang, Yang You, and Jian Zhao. 2023. Color prompting for data-free continual unsupervised domain adaptive person re-identification. *arXiv preprint arXiv:2308.10716*.

Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. 2022. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1060–1069.

Guangxing Han, Xuan Zhang, and Chongrong Li. 2023. One-shot unsupervised cross-domain person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3):1339–1351.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022.

Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. 2011. Person re-identification by descriptive and discriminative classification. In *Image Analysis: 17th Scandinavian Conference, SCIA 2011, Ystad, Sweden, May 2011. Proceedings 17*, pages 91–102. Springer.

Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. 2019. Frustratingly easy person re-identification: Generalizing person re-id in practice. *arXiv preprint arXiv:1905.03422*.

Ding Jiang and Mang Ye. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797.

Bingliang Jiao, Lingqiao Liu, Liying Gao, Guosheng Lin, Lu Yang, Shizhou Zhang, Peng Wang, and Yanning Zhang. 2022. Dynamically transformed instance normalization network for generalizable person re-identification. In *European conference on computer vision*, pages 285–301. Springer.

Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. 2020. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3143–3152.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild (2024). *URL https://llava-vl. github. io/blog/2024-05-10-llava-next-stronger-llms*.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Guang Li, Peng Liu, Xiaofan Cao, and Chunguang Liu. 2023a. Dynamic weighting network for person re-identification. *Sensors*, 23(12):5579.

Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. 2021. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6729–6738.

Hongsheng Li, Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Xia Zhao, Syed Afaq Ali Shah, and Mohammed Bennamoun. 2024. Scene graph generation: A comprehensive survey. *Neurocomputing*, 566:127052.

Huafeng Li, Yiwen Chen, Dapeng Tao, Zhengtao Yu, and Guanqiu Qi. 2020. Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification. *IEEE Transactions on Information Forensics and Security*, 16:1480–1494.

Jiachen Li and Xiaojin Gong. 2025. Unleashing the potential of pre-trained diffusion models for generalizable person re-identification. *Sensors (Basel, Switzerland)*, 25(2):552.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Siyuan Li, Li Sun, and Qingli Li. 2023c. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 1405–1413.

Wei Li and Xiaogang Wang. 2013. Locally aligned feature transforms across views. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3594–3601.

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159.

Yu Lian, Wenmin Huang, Shuang Liu, Peng Guo, Zhong Zhang, and Tariq S Durrani. 2023. Person re-identification using local relation-aware graph convolutional network. *Sensors*, 23(19):8138.

Shengcai Liao and Ling Shao. 2020. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *European conference on computer vision*, pages 456–474. Springer.

Shengcai Liao and Ling Shao. 2021. Transmatcher: Deep image matching through transformers for generalizable person re-identification. *Advances in Neural Information Processing Systems*, 34:1992–2003.

Chen Change Loy, Tao Xiang, and Shaogang Gong. 2009. Multi-camera activity correlation analysis. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995. IEEE.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0.

Hao Ni, Yuke Li, Lianli Gao, Heng Tao Shen, and Jingkuan Song. 2023. Part-aware transformer for generalizable person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11280–11289.

Hao Ni, Jingkuan Song, Xiaopeng Luo, Feng Zheng, Wen Li, and Heng Tao Shen. 2022. Meta distribution alignment for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2487–2496.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, JP Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 1912. An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst*, 32:8026.

Jinjia Peng, Song Pengpeng, Hui Li, and Huibing Wang. 2024. Refid: reciprocal frequency-aware generalizable person re-identification via decomposition and filtering. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(7):1–20.

Lei Qi, Ziang Liu, Yinghuan Shi, and Xin Geng. 2024. Generalizable metric network for cross-domain person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Lei Qi, Jiaying Shen, Jiaqi Liu, Yinghuan Shi, and Xin Geng. 2022. Label distribution learning for generalizable multi-source person re-identification. *IEEE Transactions on Information Forensics and Security*, 17:3139–3150.

Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. 2018. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7229–7238.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. 2022. A closer look at smoothness in domain adversarial training. In *International conference on machine learning*, pages 18378–18399. PMLR.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. 2021. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986.

Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. 2018. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*.

Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2019. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 719–728.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *CVPR*.

Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. 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.

Mingxing Tan and EfficientNet Le Q V. 1905. rethinking model scaling for convolutional neural networks. 2019. *arXiv preprint arXiv:1905.11946*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. 2014. Person re-identification by video ranking. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 688–703. Springer.

Zhikang Wang, Lihuo He, Xiaoguang Tu, Jian Zhao, Xinbo Gao, Shengmei Shen, and Jiashi Feng. 2021. Robust video-based person re-identification by hierarchical mining. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8179–8191.

Zhongdao Wang, Jingwei Zhang, Liang Zheng, Yixuan Liu, Yifan Sun, Yali Li, and Shengjin Wang. 2020. Cycas: Self-supervised cycle association for learning re-identifiable descriptions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 72–88. Springer.

Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88.

Suncheng Xiang, Yuzhuo Fu, Hao Chen, Wei Ran, and Ting Liu. 2020. Multi-level feature learning with attention for person re-identification. *Multimedia Tools and Applications*, 79:32079–32093.

Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3415–3424.

Haoyu Xie, Changqi Wang, Jian Zhao, Yang Liu, Jun Dan, Chong Fu, and Baigui Sun. 2024. Prcl: Probabilistic representation contrastive learning for semi-supervised semantic segmentation. *International Journal of Computer Vision*, 132(10):4343–4361.

Boqiang Xu, Jian Liang, Lingxiao He, and Zhenan Sun. 2022. Mimic embedding via adaptive aggregation: Learning generalizable person re-identification. In *European Conference on Computer Vision*, pages 372–388. Springer.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Deep metric learning for person re-identification. In *2014 22nd international conference on pattern recognition*, pages 34–39. IEEE.

Ye Yuan, Wuyang Chen, Tianlong Chen, Yang Yang, Zhou Ren, Zhangyang Wang, and Gang Hua. 2020. Calibrated domain-invariant learning for highly generalizable large scale re-identification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3589–3598.

Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Xianglong Liu, and Ziwei Liu. 2022a. Delving deep into the generalization of vision transformers under distribution shifts. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 7277–7286.

Lei Zhang, Zhipu Liu, Wensheng Zhang, and David Zhang. 2023. Style uncertainty based self-paced meta learning for generalizable person re-identification. *IEEE Transactions on Image Processing*, 32:2107–2119.

Pengyi Zhang, Huanzhang Dou, Yunlong Yu, and Xi Li. 2022b. Adaptive cross-domain learning for generalizable person re-identification. In *European conference on computer vision*, pages 215–232. Springer.

Yi-Fan Zhang, Zhang Zhang, Da Li, Zhen Jia, Liang Wang, and Tieniu Tan. 2022c. Learning domain invariant representations for generalizable person re-identification. *IEEE Transactions on Image Processing*, 32:509–523.

Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. 2021. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6277–6286.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124.

Liang Zheng, Yi Yang, and Alexander G Hauptmann. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. 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.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. 2020. Identity-guided human semantic parsing for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 346–363. Springer.