# Divide and Conquer: Hybrid Pre-training for Person Search

**Yanling Tian[1], Di Chen[1], Yunan Liu[1,2], Jian Yang[1], Shanshan Zhang[1]***

[1]PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security,
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2] School of Artificial Intelligence, Dalian Maritime University
{yl.tian, dichen, liuyunan, shanshan.zhang, csjyang}@njust.edu.cn

## Abstract

Large-scale pre-training has proven to be an effective method for improving performance across different tasks. Current person search methods use ImageNet pre-trained models for feature extraction, yet it is not an optimal solution due to the gap between the pre-training task and person search task (as a downstream task). Therefore, in this paper, we focus on pre-training for person search, which involves *detecting* and *re-identifying* individuals simultaneously. Although labeled data for person search is scarce, datasets for two sub-tasks person detection and re-identification are relatively abundant. To this end, we propose a hybrid pre-training framework specifically designed for person search using sub-task data only. It consists of a hybrid learning paradigm that handles data with different kinds of supervisions, and an intra-task alignment module that alleviates domain discrepancy under limited resources. To the best of our knowledge, this is the first work that investigates how to support full-task pre-training using sub-task data. Extensive experiments demonstrate that our pre-trained model can achieve significant improvements across diverse protocols, such as person search method, fine-tuning data, pre-training data and model backbone. For example, our model improves ResNet50 based NAE by 10.3% relative improvement w.r.t. mAP. Our code and pre-trained models are released for plug-and-play usage to the person search community (https://github.com/personsearch/PretrainPS).

## Introduction

Person search aims to localize a person and identify the person from a gallery set of real-world uncropped scene images, which can be seen as a combined pedestrian detection and re-identification (re-ID) task. Person search is an extremely difficult problem because it requires optimising these two different sub-tasks in a unified framework, and even the optimisation objectives of the two sub-tasks are inconsistent.

Most existing person search methods (Xiao et al. 2017; Li et al. 2022; Yan et al. 2022; Han et al. 2021; Chen et al. 2021; Kim et al. 2021; Li and Miao 2021; Han, Ko, and Sim 2021; Yan et al. 2021, 2023; Tian et al. 2022) use ImageNet pre-trained models (Deng et al. 2009), such as ResNet50 (He et al. 2016), as the initialization model for feature extraction. However, ImageNet pre-training, which learns classification-related knowledge, is limited in its applicability to downstream tasks, particularly when the target task is significantly different (He, Girshick, and Dollar 2019). The person search task, which includes a pedestrian detection task and a fine-grained identity classification task, requires prior knowledge in regression and there is a substantial domain gap between image classification and person re-ID tasks (Fu et al. 2021). To address this issue, task-specific pre-training models are necessary.

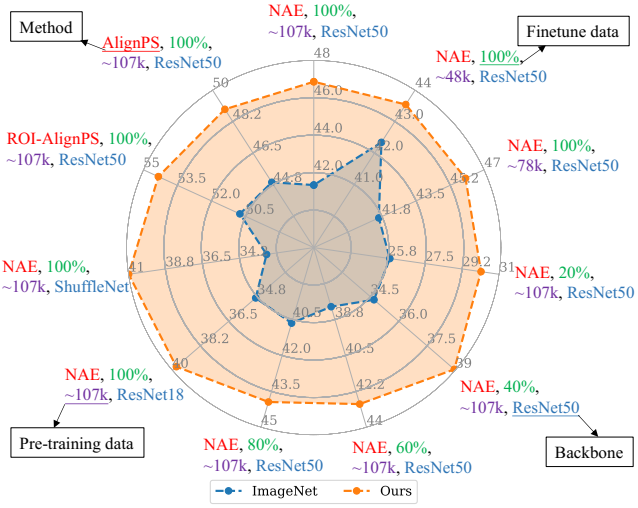



Figure 1: Performance gain using our pre-trained model instead of ImageNet model across different methods, fine-tuning data, pre-training data, backbones on PRW dataset.

Some recent works aim to design specific pre-training methods for different computer vision tasks (Fu et al. 2021, 2022; Yang et al. 2022; Bar et al. 2022). For example, DETR (Carion et al. 2020) is pre-trained using a multi-task learning approach on the COCO training dataset (118,000 images) with object annotations for both object detection and object re-localization tasks; (Carreira and Zisserman 2017) propose a new model pre-trained on the proposed Kinetics dataset with 300,000 video clips for action recognition. Similarly, we seek a suitable dataset for person search pre-training, but collecting a large-scale dataset is challenging especially when we need expensive person iden-

*Corresponding author.

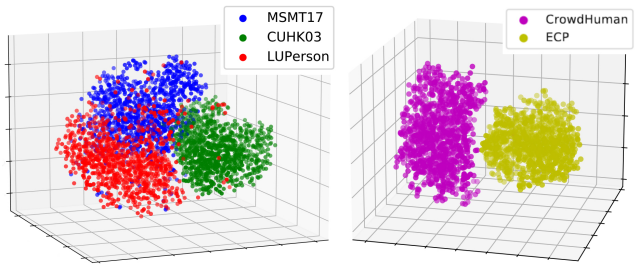
tification labels. Although unsupervised pre-training is a straightforward solution, it costs a large amount of computational resources, which are usually not affordable, for example, (Yang et al. 2022) pre-train their UP-ReID models using 168 GPU days. Aiming to minimize computational costs for pre-training person search models, we attempt to use currently available datasets with or without labels for pedestrian detection and re-ID, which are sub-tasks for person search.

Since we use sub-task datasets for pre-training, two crucial considerations must be taken into account: (1) **How to support full-task training by sub-task data?** Pedestrian detection datasets only provide person bounding box (bbox) annotations that can only facilitate training for the detection sub-task; while re-ID datasets lack background information and thus can only facilitate training for the re-ID sub-task. It is challenging to design a learning paradigm such that the entire person search model is well optimized. (2) **How to deal with domain discrepancy under limited data?** Compared to other pre-training methods that use a large amount of data, the data we can use is rather limited. We observe that samples from different datasets, such as CrowdHuman (Shao et al. 2018) and EuroCity Persons (ECP) (Braun et al. 2019), exhibit significant differences in style appearance and underlying distributions (Fig. 2). Thus, it is necessary to deal with domain discrepancy so as to enhance the generalization ability of our pre-trained models.

To facilitate pre-training on sub-task data, we present a customized pre-training method. Specifically, we propose a hybrid training paradigm so that data with different kinds of supervisions can be handled in one joint framework; also, to alleviate the negative impact brought by domain discrepancy, we propose an intra-task alignment module (IAM), which is used to align features for detection and re-ID sub-tasks separately, so that the learned representations are domain invariant. As shown in Fig. 1, our pre-trained model significantly outperforms the ImageNet pre-trained model across different protocols (*i.e.* methods, fine-tuning data, pre-training data and backbones) on the PRW dataset (Zheng et al. 2017). These results show that our approach provides stronger pre-trained models for the person search task.

In summary, our contributions can be summarized as follows:

1. Due to the lack of large-scale person search datasets, we propose to use off-the-shelf datasets of sub-tasks (pedestrian detection and re-ID task) for person search pre-training. To the best of our knowledge, this is the first work that investigates how to support full-task pre-training using sub-task data for person search. We believe our attempt is inspiring and encouraging for future work.
2. We propose a novel pre-training method specific for person search. It consists of a hybrid learning paradigm that handles data with different kinds of supervisions, and an intra-task alignment module that alleviates domain discrepancy under limited resources.
3. We provide analyses showing that our pre-training method is generalizable across different backbones and our pre-trained models benefit different methods. Our pre-trained models are more effective than the ImageNet ones and thus are expected to essentially contribute to the person search community. Benefited from our pre-training method, we establish a new state of the art on the CUHK-SYSU (Xiao et al. 2017) and PRW (Zheng et al. 2017) benchmarks.




(a) re-ID datasets (b) detection datasets

Figure 2: t-SNE visualization of features with ResNet50 pre-trained on ImageNet from 1,000 random samples in different detection and re-ID datasets.

## Related Work

Since this paper develops pre-trained models specifically promoting person search, we thereby review relevant works from the above two aspects.

### Person Search

Due to its wide application prospects, person search has made rapid progress in recent years. Some works (Wang et al. 2020; Yao and Xu 2020; Dong et al. 2020b; Han et al. 2019; Chen et al. 2020b; Lan, Zhu, and Gong 2018a) optimize the pedestrian detection and re-ID tasks in person search separately (two-stage methods), while others (Xiao et al. 2017; Li et al. 2022; Cao et al. 2022; Yan et al. 2022; Yu et al. 2022; Han et al. 2021; Chen et al. 2021; Kim et al. 2021; Li and Miao 2021; Han, Ko, and Sim 2021; Yan et al. 2021, 2023) regard two sub-tasks as a whole and jointly optimize them in an end-to-end manner (one-stage methods). Generally, existing person search methods can be roughly divided into two groups: fully supervised and weakly supervised methods. **(1)** ***Fully supervised.*** Given human bbox and their corresponding identity information, this group of methods allows for direct model training. As a pioneering method for person search, (Xiao et al. 2017) propose a joint framework that incorporates re-ID layers on top of Faster-RCNN (Ren et al. 2015) detector. (Chen et al. 2021) introduce a norm-aware embedding method (NAE) to accommodate the conflicting optimization objectives of detection and re-ID. (Yan et al. 2021) construct an anchor-free framework to address the misalignment in different levels (including scales, regions, and tasks). To improve performance, some recent methods (Cao et al. 2022; Yu et al. 2022) utilize transformer to learn more discriminative feature representations. **(2)** ***Weakly supervised.*** Considering that obtaining identity information is more difficult, this group of methods only uses annotations of bbox for training. (Han et al. 2021)

propose a region siamese network for recognition in the absence of identity annotations. (Yan et al. 2022) propose a weakly supervised person search method by leveraging context clues of detection, memory and scene in unconstrained natural images.

In this paper, we study person search from a novel perspective, making full use of unlabeled and labeled data of sub-tasks to develop powerful pre-trained models for person search.

## Vision Model pre-training

Benefiting from the strong visual knowledge distributed in ImageNet (Deng et al. 2009), fine-tuning a pre-trained model with a small amount of task-specific data can perform well on downstream tasks. This triggers the first wave of exploring pre-trained models in the era of deep learning.

The early efforts of pre-training are mainly achieved in a supervised manner. By applying ResNet (He et al. 2016) pre-trained on ImageNet as the backbone, various vision tasks (*e.g.* image classification and segmentation) have been quickly advanced. In comparison with supervised pre-training, self-supervised pre-training allows for huge advances. Some methods (Fu et al. 2021, 2022; Chen et al. 2018a) construct positive pairs with data augmentation, and obtain pre-trained models via contrastive learning (Chen et al. 2020d). (Bar et al. 2022) propose a self-supervised method that pre-trains the entire object detection network, including the object localization and embedding components. (He et al. 2022) adopt an asymmetric encoder-decoder network and show that scalable vision learners can be obtained simply by reconstructing the missing pixels. (Wei et al. 2022) use histograms of oriented gradients to learn abundant visual knowledge for pre-training video models. (Fu et al. 2021) study the key factors (*i.e.* data augmentation and contrastive loss) to improve the generalization ability of learned re-ID features. Based on the contrastive learning pipeline, (Yang et al. 2022) propose an unsupervised framework to learn the fine-grained re-ID features. Within a contrastive learning framework, (Shuai et al. 2022) attempt to learn person similarity without using manually labeled identity annotations.

Compared to individual task pre-training, it is much more challenging to pre-train for a hybrid task like person search, due to the lack of large-scale datasets. In this paper, we investigate how to support full-task pre-training using sub-task data, which has not been studied by previous works and our findings are expected to be inspiring and encouraging for future works.

# Method

In this section, we begin with an overview of our pre-training framework, followed by an explanation of the proposed hybrid learning approach. Due to the domain discrepancy in the hybrid learning, we introduce our simple intra-task alignment module (IAM) in detail at last.

## Overview

In this paper, we propose a novel pre-training method, which distills specific knowledge from data of two sub-tasks to promote person search. In Fig. 3, we show the pipeline of our pre-training method, which is trained using a hybrid learning approach in an end-to-end manner. Following Simsiam (Chen and He 2021), our framework consists of a shared person search model, a prediction multilayer perceptron (MLP), a data unification module and IAMs. As shown in Fig. 3, we randomly sample images $x^D$ and $x^R$ from detection and re-ID datasets and perform data unification on them to handle different data types, where $x^D = \{x^{D_i} | D_i \in \{D_1, ..., D_M\}\}$ and $x^R = \{x^{R_j} | R_j \in \{R_1, ..., R_N\}\}$. Initially, we apply two data augmentations to each image, resulting in two different views of the same image. Since we use sub-task datasets to train a person search model, specific flows are designed for different kinds of data: (1) For $x$ from any dataset, each view goes through the entire person search model for detection task training and re-ID task training in a fully supervised way. (2) For $x$ from datasets without person identities, *i.e.* unlabeled re-ID datasets and detection datasets, it is used for re-ID task training in a self-supervised way, *i.e.* two views are sent to the framework for contrastive learning. Moreover, two IAMs for detection and re-ID respectively, are developed to alleviate domain discrepancy by aligning features across different datasets.

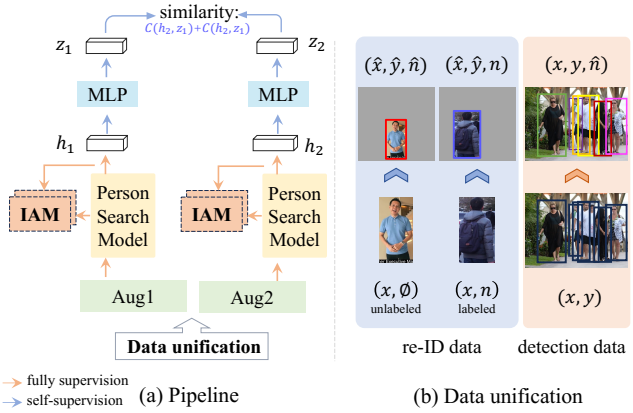



Figure 3: The pipeline of our pre-training method. MLP refers to a prediction multilayer perceptron. IAM is used to alleviate domain discrepancy across different datasets. The right figure is the data unification operation to unify different data types for a unified interface. Black bboxes denote the people with no identity while colorful bboxes refer to persons with different identities.

## Hybrid Learning

In a dilemma of lacking large-scale person search datasets, we use sub-task datasets for training. Each dataset only provides labels for each sub-task, or even no labels at all like the unlabeled re-ID datasets. In order to handle different types of data in a unified framework and provide a simplified training pipeline, we propose a hybrid learning approach that combines self-supervised and fully-supervised learning to assist the unified framework in gathering knowledge from differ-

ent datasets. The loss function for the entire pre-training procedure can be expressed as:

$$L = L_{ps} + \eta L_{con} + \lambda L_{adv}, \tag{1}$$

where $\lambda$ and $\eta$ are hyper-parameters; $L_{ps}$ (Eq. 4), $L_{con}$ (Eq. 6) and $L_{adv}$ (Eq. 7) are introduced in the following.

**Data unification** The schematic diagram of data unification is shown in Fig. 3 (b).

For pedestrian detection datasets, $D_i=(x^{D_i}, y^{D_i})_{i=1}^{N_{D_i}}$ only provide position information of each person, without identity. $x$ and $y$ denote the image and person bbox annotations respectively; and $N_{D_i}$ is the total number of image samples in the corresponding dataset. To construct a unified interface and simplified pipeline, we regard each person in $D_i$ as different persons. Therefore, we have the detection dataset $D_i=(x^{D_i}, y^{D_i}, \hat{n}_i^{D_i})_{i=1}^{N_{D_i}}$, $\hat{n}$ is the new identity annotation.

For the re-ID datasets, the images form datasets are cropped from entire images, resulting in the absence of contextual background information. We obtain labeled re-ID dataset $R_j=(x_j^{R_j}, n_j^{R_j})_{j=1}^{N_{R_j}}$ and the unlabeled re-ID dataset $R_j^{un}=(x_j^{R_j^{un}}, )_{j=1}^{N_{R_j^{un}}}$ based on the presence of identity annotation. $x$ and $n$ denote the image and identity annotation. $N_{R_j}$, $N_{R_j^{un}}$ are the number of image samples in corresponding datasets. To construct a unified interface and simplified pipeline, similar to detection datasets, we also regard each person as a new individual. In addition, we randomly put the re-ID image on a canvas of varying proportions to the size of the image and resize to a fixed size. We refer to this operation as "expand_resize". Therefore, we have the labeled re-ID dataset $R_j=(\hat{x}_j^{R_j}, \hat{y}_j^{R_j}, n_j^{R_j})_{j=1}^{N_{R_j}}$ and unlabeled re-ID dataset $R_j^{un}=(\hat{x}_j^{R_j^{un}}, \hat{y}_j^{R_j^{un}}, \hat{n}_j^{R_j^{un}})_{j=1}^{N_{R_j^{un}}}$, where $\hat{x}$ denotes the canvas containing re-ID samples, $\hat{n}$ is new identity annotation, $\hat{y}$ is the location of each person on its canvas.

**Fully-supervised learning** After data unification, we can perform person search pre-training in a fully-supervised way. Given an arbitrary image $x$ and its person bbox annotations $y$, we optimize the encoder and detection head in person search model by the following detection loss function:

$$L_{det} = \sum_{i=0}^{J} L_{rpn}(x_i, y_i) + L_{det-head}(x_i, y_i), \tag{2}$$

where $L_{rpn}$ indicates the loss for the RPN network (Ren et al. 2015); $L_{det-head}$ is the loss for the detection head. $J$ is the image number of all sub-task datasets. With the image $x$ and its identity annotation $n$, we optimize the encoder and the re-ID head via the following re-ID loss function:

$$L_{reid} = \sum_{i=0}^{J} L_{oim}(x_i, n_i), \tag{3}$$

where $L_{oim}$ is identical to that used by NAE (Xiao et al. 2017). Please note person samples from $x^{D_i}$ and $x^{R^{un}}$ serve as different identities in $L_{oim}$ (Xiao et al. 2017), so that person representation learning can benefit from more unlabeled samples.

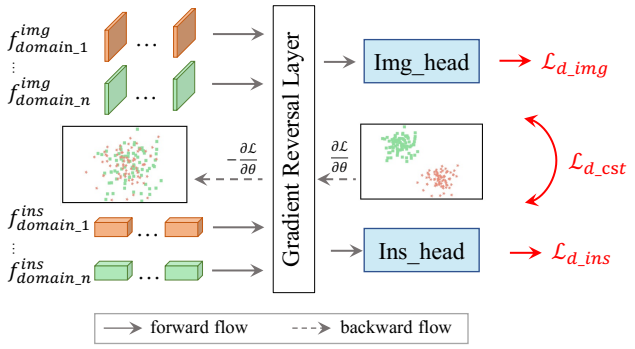



Figure 4: Structure of intra-task alignment module (IAM). Gradient reversal layer (Ganin and Lempitsky 2015) automatically reverses the gradient during backward propagation to achieve feature alignment across different datasets (domains) with the heads. In the backward flow, the colorful points refer to different features from different domains.

Therefore, the total loss function for fully-supervised learning can be written as:

$$L_{ps} = L_{det} + L_{reid}. \tag{4}$$

**Self-supervised learning** In order to better distill re-ID knowledge from images without ID labels, we perform self-supervised learning for the re-ID sub-task. Given an arbitrary image $x$ sampled from $\{D_1, ..., D_M\}$ or $R^{un}$, we use two different augmentations to generate different views $Aug_1(x)$ and $Aug_2(x)$. Each view $Aug_i(x)$ goes through the entire person search model to obtain deep feature representations, denoted by $h_i = P(Aug_i(x))$, where $P$ is the entire person search model. And then $h_i$ is transformed to $z_i$ by the prediction MLP head. To maintain the consistency of predictions from two views, we minimize the negative cosine similarity as follows:

$$C(h_2, z_1) = -\sum_{k=0}^{K} \frac{h_2^k}{||h_2^k||_2} \cdot \frac{z_1^k}{||z_1^k||_2}, \tag{5}$$

where $||\cdot||_2$ is $l_2$ norm. $K$ refers to the image number of detection datasets and unlabeled re-ID dataset. Inspired by a recent work on contrastive learning (Chen and He 2021), we further adopt a symmetrized loss in self-supervised learning as,

$$L_{con}(h, z) = \frac{1}{2} C(sg(h_2), z_1) + \frac{1}{2} C(sg(h_1), z_2), \tag{6}$$

where $sg$ denotes the operation of stop-gradient (Chen and He 2021). In this way, $P$ receives no gradient from $z_i$ in order to alleviate collapse solutions in siamese network during optimization.

## Intra-task Alignment Module

There are differences in data distribution among datasets of the same task (Fig. 2), making it difficult to ensure effective distillation of task-related knowledge. To solve this problem, we present an IAM, which performs adversarial-based feature alignment at both instance and image levels. As shown

in Fig. 3, we use two separate IAMs (*i.e.* detection IAM and re-ID IAM) to reduce the discrepancy for pedestrian detection and re-ID datasets, respectively. The reason for such design is that pedestrian detection and re-ID conflict with each other in terms of features. Pedestrian detection task is concerned with identifying shared features among individuals, whereas re-ID aims to capture the distinct features and specific details of each identity.

The architecture of detection IAM and re-ID IAM is identical. As shown in Fig. 4, we use features at both image level $f^{img}$ and instance level $f^{ins}$ as the input of IAM. For detection IAM, $f^{img}$ refers to the output feature maps of the encoder in the person search model, while $f^{ins}$ is the feature vectors from the detection head of the person search model. For re-ID IAM, $f^{img}$ is the feature maps extracted by the ROI-Align operation and $f^{ins}$ represents the feature vectors from the re-ID head in the person search model. Subsequently, these features are passed through several cascaded convolutional layers, followed by two fully connected (FC) layers termed as Img_head and Ins_head respectively. Finally, the predictions from two heads are used to perform domain classification based on $f^{img}$ and $f^{ins}$, which is optimized by the following loss function:

$$\begin{aligned} L_{adv}=&L_{d\_img}(x^D, x^R, \overline{y}) + L_{d\_ins}(x^D, x^R, \overline{y}) \\ &+ L_{d\_cst}(x^D, x^R), \end{aligned} \tag{7}$$

where $\overline{y}$ indicates the domain label (*i.e.* to distinguish which dataset an image belongs to); $L_{d\_img}$ and $L_{d\_ins}$ are multi-class cross-entropy loss that applied to the image level and instance level respectively. We note that maintaining the consistency between the domain classifiers at both levels is helpful to improve performance (Chen et al. 2018b) (from which we observe a 0.3pp increase w.r.t. mAP.). To this end, we introduce a consistency regularizer $L_{d\_cst}$, following (Chen et al. 2018b). After applying our IAM, we obtain domain-invariant features that are agnostic to the specific data domain.

# Experiments

In this section, we first introduce benchmark datasets and evaluation metrics. Then, a series of experimental analyses are conducted. Finally, we compare with state-of-the-art methods on multiple benchmarks. Implementation details are provided in our code.

## Datasets and Evaluation Metrics

**Datasets for pre-training.** We use two relatively large person detection datasets, *i.e.* CrowdHuman (Shao et al. 2018) and EuroCity Persons (ECP) (Braun et al. 2019) datasets. In addition, we use two relatively common re-ID datasets including MSMT17 (Wei et al. 2018) and CUHK03 (Li et al. 2014), and one unlabeled large re-ID dataset, LUPerson (Fu et al. 2021) for pre-training.

**Person search datasets.** The two most commonly used datasets for person search are PRW (Zheng et al. 2017) and CUHK-SYSU (Xiao et al. 2017). **CUHK-SYSU** is a large-scale person search dataset, providing 18,184 images and 8,432 individuals, with some images sourced from movie snapshots and others from street/city scenes. The training set consists of 11,206 images and 5,532 identities, while the testing data contains 6,978 gallery images with 2,900 query persons. Instead of using the entire testing set as a gallery, we follow the standard protocols with gallery sizes ranging from 50 to 4,000. We use the default gallery size of 100 in our experiments unless otherwise specified. **PRW** includes 11,816 images captured with six cameras on a university campus. It provides 5,134 training images with 482 identities while the testing set includes 2,057 different query persons and 6,112 gallery images. We use all gallery images as the search space for each query person. In addition to these two standard datasets, there is a new dataset, **PoseTrack21** (Doering et al. 2022), that can be used for person search. It consists of 42,861 training images with 5,474 different individuals, and 19,935 gallery images with 1,313 query individuals. Unlike the above datasets, the queries in PoseTrack21 may contain multiple individuals in cases of occlusion.

| Backbone | mAP | Top-1 |
|---|---|---|
| ResNet50 | 42.67→47.08 (↑10.3%) | 81.33→84.00 |
| ResNet18 | 35.87→39.79 (↑10.9%) | 77.43→79.76 |
| ShuffleNet | 34.29→40.98 (↑19.5%) | 78.07→80.74 |

Table 1: Performance gain of NAE (Chen et al. 2020c) with different pre-trained backbones. A→B: A refers to the performance using ImageNet pre-trained model, while B is that using our pre-trained model. The number after ↑ is relative improvement.

**Evaluation Metrics.** mAP and Top-$k$ cumulative matching characteristics are the performance metrics for person search. The mAP metric refers to the accuracy and matching rate of searching a query person from the gallery images. The Top-$k$ score reflects the percentage of queries for which at least one of the $k$ most similar proposals succeeds in the re-ID matching step.

## Analysis

In this section, we conduct a series of analyses on our method. Unless otherwise specified, our pre-training method is trained on five datasets with a total of 106,784 images, including CrowdHuman, ECP, MSMT17, CUHK03 and LUPerson30k datasets. All experiments in this subsection use PRW as the target dataset for verification unless otherwise specified.

**Generalizable to different backbones** We apply our pre-training method on top of different backbones: ResNet18 (He et al. 2016), ResNet50 (He et al. 2016), ShuffleNet (Zhang et al. 2018), and provide three pre-trained models to a method NAE. As shown in Tab. 1, compared to ImageNet pre-trained models, our pre-trained models achieve more than 10% relative improvements w.r.t. mAP across three different backbones, and especially the gain is up to 19.5% for ShuffleNet. These consistently significant

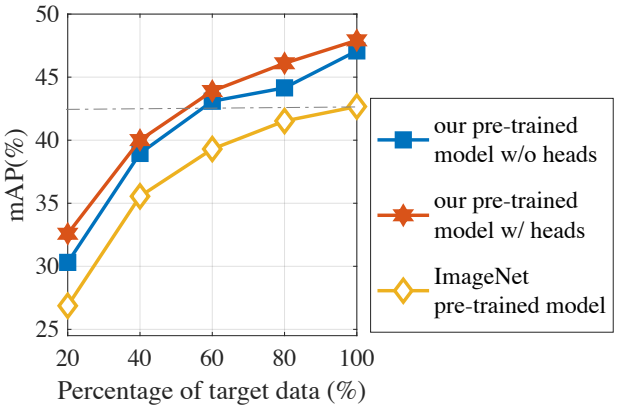



Figure 5: Performance comparison for NAE (Chen et al. 2020c) with different pre-trained models on PRW dataset. By using only 60% of target data, our pre-trained model outperforms the ImageNet pre-trained model trained on all annotations.

| Method | mAP | Top-1 |
|---|---|---|
| NAE | 42.67→47.08 (↑4.41) | 81.33→84.00 (↑2.67) |
| AlignPS | 45.90→49.14 (↑3.24) | 81.90→83.56 (↑1.66) |
| ROI-AlignPS | 51.84→54.47 (↑2.63) | 85.48→86.56 (↑1.08) |

Table 2: Performance gain of different person search methods using our pre-trained model instead of ImageNet model. A→B: A refers to the performance using ImageNet pre-trained model, while B is that using our pre-trained model.

improvements demonstrate our pre-training approach well generalizes to various backbones.

**Beneficial for different methods and datasets** We also feed our pre-trained model (ResNet50) to different person search methods for initialization, including an anchor-based method (NAE (Chen et al. 2020c)), an anchor-free method (AlignPS (Yan et al. 2021)) and ROI-AlignPS (Yan et al. 2023) (mixing NAE and AlignPS). As shown in Tab. 2, all methods improve significantly over its counterpart using ImageNet pre-trained model, w.r.t. both mAP and Top-1 accuracy. It is notable that NAE surpasses AlignPS when switched to our pre-trained model. Even for a very strong method ROI-AlignPS, the improvement is still impressive, *i.e.* ∼3pp w.r.t. mAP. These results indicate that our pre-trained models are beneficial to various methods.

Due to ROI-AlignPS (Yan et al. 2023) is the current available state-of-the-art one-stage method, we use it as baseline and fine-tune it on different datasets. As shown in Tab. 3, the results with our pre-trained models are better than those with ImageNet pre-trained models.

**Requiring less target training data** We observe the performance of NAE when less target training data is used for fine-tuning. From Fig. 5, we have the following observations: (1) When using our pre-trained model, it outperforms its counterpart using the ImageNet pre-trained model by a large margin when different amounts of target data are used for fine-tuning. (2) Our pre-trained NAE surpasses its ImageNet pre-trained counterpart, when using only 60% of the target data. (3) When we initialize the entire model (including both backbones and heads) with our pre-trained model, the performance gain is larger than that solely initializes backbone, parameters, which is consistent with (Bar et al. 2022). These results demonstrate our pre-training model provides a better initialization so that the person search method requires much less target data for fine-tuning to achieve comparable performance.

| Dataset | mAP | Top-1 |
|---|---|---|
| PRW | 51.84→54.47 | 85.48→86.56 |
| CUHK-SYSU | 95.33→95.38 | 95.76→96.03 |
| PoseTrack21 | 59.21→63.62 | 82.10→87.13 |

Table 3: Performance gain of ROI-AlignPS (Yan et al. 2023) on different datasets using our pre-trained model instead of ImageNet model. A→B: A refers to the performance using ImageNet pre-trained model, while B is that using our pre-trained model.

| Operation | mAP | Top-1 |
|---|---|---|
| random_paste | 46.49 | 83.36 |
| expand_resize | **47.08** | **84.00** |

Table 4: NAE performance (Chen et al. 2020c) of using different re-ID data operation in data unification. "random_paste": paste a re-ID image randomly on a fixed size canvas. "expand_resize": put re-ID image on a canvas of different ratios of original image and resize to a fixed size.

**Effect of re-ID data operation in data unification** Multi-scale matching problem is an underlying challenge of person search (Lan, Zhu, and Gong 2018b). We obtain a scale-invariant feature through pre-training by placing re-ID images on canvases of different proportions and resizing them to a fixed size, which helps alleviate the matching problem. We refer to our approach as "expand_resize" and the operation of randomly pasting re-ID images onto a fixed-size canvas as "random_paste". As shown in Tab. 4, we validate the effectiveness of "expand_resize".

**Impact of our IAM** We further investigate the effect of our proposed IAM module, which alleviates domain discrepancy by feature alignment. From Fig. 6, we can see that: (1) When we use more datasets for pre-training, the performance gain becomes smaller, from 2.12pp for 4 datasets to 0.6pp for 5 datasets. (2) By employing our IAM, the performance gain grows at different number of datasets.

It is possible to perform both intra-task and inter-task alignment to alleviate domain discrepancy. We study their effects by adding them one by one. From Tab. 5, we can see that intra-task alignment achieves 47.08pp w.r.t. mAP, yet

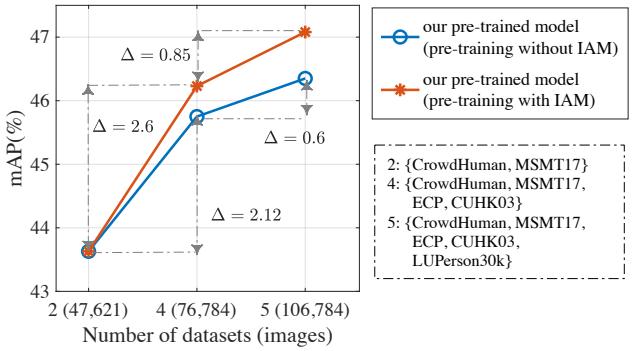



Figure 6: Effect of IAM with increasing pre-training datasets.

| intra-task AMs (IAMs) | inter-task AM | mAP | Top-1 |
|---|---|---|---|
| ✓ | | **47.08** | **84.00** |
| ✓ | ✓ | 46.53 | 83.16 |

Table 5: Effects of intra-task alignment module (IAM) and inter-task alignment module.

inter-task alignment shows a negative effect by dropping the performance slightly. The reason is that different tasks are of different optimization goals, and thus it is unfavorable to align the features across two tasks. Based on the above observations, we only use intra-task alignment in our method.

## Comparison with SOTA Methods

In this section, the backbones of all the methods are the ResNet50. We compare our method with the one-stage methods OIM (Xiao et al. 2017), NPSM (Liu et al. 2017), RCAA (Chang et al. 2018), CTXG (Yan et al. 2019), QEEPS (Munjal et al. 2019), HOIM (Chen et al. 2020a), BINet (Dong et al. 2020a), NAE (Chen et al. 2020c), PGA (Kim et al. 2021), SeqNet (Li and Miao 2021), AGWF (Han, Ko, and Sim 2021), AlignPS (Yan et al. 2021), PSTR (Cao et al. 2022), COAT (Yu et al. 2022), ROI-AlignPS (Yan et al. 2023) and two-stage methods DPM+IDE (Zheng et al. 2017), CNN+CLSA (Lan, Zhu, and Gong 2018a), FPN+RDLR (Han et al. 2019), IGPN (Dong et al. 2020b), OR (Yao and Xu 2020), TCTS (Wang et al. 2020) on the common CUHK-SYSU dataset (Xiao et al. 2017) and PRW dataset (Zheng et al. 2017).

**Results on CUHK-SYSU and PRW datasets**. As shown in Tab. 6, we achieve 95.4pp and 96.0pp w.r.t. mAP and Top-1 scores respectively, establishing a state-of-the-art performance on CUHK-SYSU dataset. Due to the large size of the CUHK-SYSU dataset and the relatively small number of gallery samples for each query, it is relatively easy to achieve saturation performance on this dataset. As a result, our method only surpasses the baseline (*i.e.* ROI-AlignPS†) by 0.1pp. In contrast, the PRW dataset has less training data and a larger gallery, resulting in a significant performance degradation. However, our method achieves 54.5pp w.r.t. mAP on the PRW dataset, obtaining an improvement of 2.7pp over the baseline ROI-AlignPS†.

| | Methods | CUHK-SYSU | | PRW | |
|---|---|---|---|---|---|
| | | mAP | Top-1 | mAP | Top-1 |
| two-stage | DPM+IDE | - | - | 20.5 | 48.3 |
| | CNN+CLSA | 87.2 | 88.5 | 38.7 | 65.0 |
| | FPN+RDLR | 93.0 | 94.2 | 42.9 | 70.2 |
| | IGPN | 90.3 | 91.4 | 47.2 | 87.0 |
| | OR | 92.3 | 93.8 | **52.3** | 71.5 |
| | TCTS | **93.9** | **95.1** | 46.8 | **87.5** |
| one-stage | OIM | 75.5 | 78.7 | 21.3 | 49.4 |
| | NPSM | 77.9 | 81.2 | 24.2 | 53.1 |
| | RCAA | 79.3 | 81.3 | - | - |
| | CTXG | 84.1 | 86.5 | 33.4 | 73.6 |
| | QEEPS | 88.9 | 89.1 | 37.1 | 76.7 |
| | HOIM | 89.7 | 90.8 | 39.8 | 80.4 |
| | BINet | 90.0 | 90.7 | 45.3 | 81.7 |
| | NAE† | 91.5 | 92.4 | 42.7 | 81.3 |
| | PGA | 92.3 | 94.7 | 44.2 | 85.2 |
| | SeqNet | 93.8 | 94.6 | 46.7 | 83.4 |
| | AGWF | 93.3 | 94.2 | 53.3 | 87.7 |
| | AlignPS | 93.1 | 93.4 | 45.9 | 81.9 |
| | PSTR | 93.5 | 95.0 | 49.5 | **87.8** |
| | COAT | 94.2 | 94.7 | 45.9 | 81.9 |
| | ROI-AlignPS† | 95.3 | 95.8 | 51.8 | 85.5 |
| | ROI-AlignPS w/ ours | **95.4** | **96.0** | **54.5** | 87.6 |

Table 6: Comparison with state-of-the-art methods on CUHK-SYSU and PRW datasets. Best results are bold and the second results are underlined. † refers to our re-implementation.

## Conclusion

In this work, we focus on designing a specific pre-training method for the person search task. Considering the lack of large-scale person search datasets, we employ its sub-task (pedestrian detection and re-ID) datasets for pre-training and propose a unified framework to handle those datasets with and without labels, as well as large domain gaps. Specifically, our proposed method consists of a hybrid learning paradigm that handles data with different kinds of supervisions, and an intra-task alignment module that alleviates domain discrepancy under limited resources. We validate the effectiveness of our approach by providing insightful analyses from different perspectives. Additionally, we provide better pre-trained models than ImageNet ones for the person search community, which can be simply loaded by various methods for initialization to achieve higher performance. We use 106,784 images (20× larger than the data in the PRW dataset) for pre-training. As shown in Fig. 6, the performance of our approach continues to improve with an increasing number of pre-training images. We will use more data to pre-train a more powerful model in the future.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62322602, 62172225, 62372077), Natural Science Foundation of Jiangsu Province, China (Grant No. BK20230033), CAAI-Huawei MindSpore Open Fund and the China Postdoctoral Science Foundation (Grant No. 2022M720624).

## References

Bar, A.; Wang, X.; Kantorov, V.; Reed, C. J.; Herzig, R.; Chechik, G.; Rohrbach, A.; Darrell, T.; and Globerson, A. 2022. DETReg: Unsupervised Pretraining with Region Priors for Object Detection. In *Computer Vision and Pattern Recognition*, 14605–14615.

Braun, M.; Krebs, S.; Flohr, F. B.; and Gavrila, D. M. 2019. EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8): 1844–1861.

Cao, J.; Pang, Y.; Anwer, R. M.; Cholakkal, H.; Xie, J.; Shah, M.; and Khan, F. S. 2022. PSTR: End-to-End One-Step Person Search With Transformers. In *Computer Vision and Pattern Recognition*, 9458–9467.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Computer Vision and Pattern Recognition*, 6299–6308.

Chang, X.; Huang, P.-Y.; Shen, Y.-D.; Liang, X.; Yang, Y.; and Hauptmann, A. G. 2018. RCAA: Relational context-aware agents for person search. In *European conference on computer vision*, 84–100.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Schiele, B. 2020a. Hierarchical online instance matching for person search. In *AAAI Conference on Artificial Intelligence*, volume 34, 10518–10525.

Chen, D.; Zhang, S.; Ouyang, W.; Yang, J.; and Tai, Y. 2020b. Person search by separated modeling and a mask-guided two-stream CNN model. *IEEE Transactions on Image Processing*, 29: 4669–4682.

Chen, D.; Zhang, S.; Yang, J.; and Schiele, B. 2020c. Norm-aware embedding for efficient person search. In *Computer Vision and Pattern Recognition*, 12615–12624.

Chen, D.; Zhang, S.; Yang, J.; and Shiele, B. 2021. Norm-Aware Embedding for Efficient Person Search and Tracking. *International Journal of Computer Vision*, 129(11): 3154–3168.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2018a. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020d. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Computer Vision and Pattern Recognition*, 15750–15758.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Gool, L. V. 2018b. Domain adaptive Faster R-CNN for object detection in the wild. In *Computer Vision and Pattern Recognition*, 3339–3348.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Li, F.-F. 2009. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, 248–255.

Doering, A.; Chen, D.; Zhang, S.; Schiele, B.; and Gall, J. 2022. PoseTrack21: A Dataset for Person Search, Multi-Object Tracking and Multi-Person Pose Tracking. In *Computer Vision and Pattern Recognition*, 20963–20972.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020a. Bi-directional interaction network for person search. In *Computer Vision and Pattern Recognition*, 2839–2848.

Dong, W.; Zhang, Z.; Song, C.; and Tan, T. 2020b. Instance guided proposal network for person search. In *Computer Vision and Pattern Recognition*, 2585–2594.

Fu, D.; Chen, D.; Bao, J.; Yang, H.; Yuan, L.; Zhang, L.; Li, H.; and Chen, D. 2021. Unsupervised Pre-training for Person Re-identification. In *Computer Vision and Pattern Recognition*, 14750–14759.

Fu, D.; Chen, D.; Yang, H.; Bao, J.; Yuan, L.; Zhang, L.; Li, H.; Wen, F.; and Chen, D. 2022. Large-Scale Pre-training for Person Re-identification with Noisy Labels. In *Computer Vision and Pattern Recognition*, 2476–2486.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 1180–1189.

Han, B.-J.; Ko, K.; and Sim, J.-Y. 2021. End-to-end trainable trident person search network using adaptive gradient propagation. In *International Conference on Computer Vision*, 925–933.

Han, C.; Su, K.; Yu, D.; Yuan, Z.; Gao, C.; Sang, N.; Yang, Y.; and Wang, C. 2021. Weakly supervised person search with region siamese networks. In *International Conference on Computer Vision*, 12006–12015.

Han, C.; Ye, J.; Zhong, Y.; Tan, X.; Zhang, C.; Gao, C.; and Sang, N. 2019. Re-id driven localization refinement for person search. In *International Conference on Computer Vision*, 9814–9823.

He, K.; Girshick, R.; and Dollar, P. 2019. Rethinking imagenet pre-training. In *Computer Vision and Pattern Recognition*, 4918–4927.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 770–778.

He, K. H.; Chen, X.; Xie, S.; Li, Y.; Dolla'r, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *Computer Vision and Pattern Recognition*, 16000–16009.

Kim, H.; Joung, S.; Kim, I.-J.; and Sohn, K. 2021. Prototype-guided saliency feature learning for person

search. In *Computer Vision and Pattern Recognition*, 4865–4874.

Lan, X.; Zhu, X.; and Gong, S. 2018a. Person search by multi-scale matching. In *European conference on computer vision*, 536–552.

Lan, X.; Zhu, X.; and Gong, S. 2018b. Person search by multi-scale matching. In *International Conference on Computer Vision*, 536–552.

Li, J.; Yan, Y.; Wang, G.; Yu, F.; Jia, Q.; and Ding, S. 2022. Domain adaptive person search. In *European conference on computer vision*, 302–318.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. DeepReID: Deep filter pairing neural network for person re-identification. In *Computer Vision and Pattern Recognition*, 152–159.

Li, Z.; and Miao, D. 2021. Sequential end-to-end network for efficient person search. In *AAAI Conference on Artificial Intelligence*, volume 35, 2011–2019.

Liu, H.; Feng, J.; Jie, Z.; Jayashree, K.; Zhao, B.; Qi, M.; Jiang, J.; and Yan, S. 2017. Neural person search machines. In *International Conference on Computer Vision*, 493–501.

Munjal, B.; Amin, S.; Tombari, F.; and Galasso, F. 2019. Query-guided end-to-end person search. In *Computer Vision and Pattern Recognition*, 811–820.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. volume 28.

Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; and Sun, J. 2018. CrowdHuman: A Benchmark for Detecting Human in a Crowd. *arXiv preprint arXiv:1805.00123*.

Shuai, B.; Li, X.; Kundu, K.; and Tighe, J. 2022. Id-Free Person Similarity Learning. In *Computer Vision and Pattern Recognition*, 14689–14699.

Tian, Y.; Chen, D.; Liu, Y.; Zhang, S.; and Yang, J. 2022. Grouped Adaptive Loss Weighting for Person Search. In *ACM International Conference on Multimedia*, 6774–6782.

Wang, C.; Ma, B.; Chang, H.; Shan, S.; and Chen, X. 2020. TCTS: A task-consistent two-stage framework for person search. In *Computer Vision and Pattern Recognition*, 11952–11961.

Wei, C.; Fan, H.; Xie, S.; Wu, C.-Y.; Yuille, A.; and Feichtenhofer, C. 2022. Masked Feature Prediction for Self-Supervised Visual Pre-Training. In *Computer Vision and Pattern Recognition*, 14668–14678.

Wei, L.; Shiliang, Z.; Gao, W.; and Tian, Q. 2018. Person transfer GAN to bridge domain gap for person re-identification. In *Computer Vision and Pattern Recognition*, 79–88.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *Computer Vision and Pattern Recognition*, 3415–3424.

Yan, Y.; Li, J.; Liao, S.; Qin, J.; Ni, B.; Lu, K.; and Yang, X. 2022. Exploring visual context for weakly supervised person search. In *AAAI Conference on Artificial Intelligence*, volume 36, 3027–3035.

Yan, Y.; Li, J.; Qin, J.; Bai, S.; Liao, S.; Liu, L.; Zhu, F.; and Shao, L. 2021. Anchor-free person search. In *Computer Vision and Pattern Recognition*, 7690–7699.

Yan, Y.; Li, J.; Qin, J.; Liao, S.; and Yang, X. 2023. Efficient Person Search: An Anchor-Free Approach. *International Journal of Computer Vision*, 131(7): 1642–1661.

Yan, Y.; Zhang, Q.; Ni, B.; Zhang, W.; Xu, M.; and Yang, X. 2019. Learning context graph for person search. In *Computer Vision and Pattern Recognition*, 2158–2167.

Yang, Z.; Jin, X.; Zheng, K.; and Zhao, F. 2022. Unleashing Potential of Unsupervised Pre-Training with Intra-Identity Regularization for Person Re-Identification. In *Computer Vision and Pattern Recognition*, 14298–14307.

Yao, H.; and Xu, C. 2020. Joint person objectness and repulsion for person search. *IEEE Transactions on Image Processing*, 30: 685–696.

Yu, R.; Du, D.; LaLonde, R.; Davila, D.; Funk, C.; Hoogs, A.; and Clipp, B. 2022. Cascade Transformers for End-to-End Person Search. In *Computer Vision and Pattern Recognition*, 7267–7276.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Computer Vision and Pattern Recognition*, 6848–6856.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *Computer Vision and Pattern Recognition*, 1367–1376.