Thesis Reproduction of Seqnet: a pedestrian search method

Qipeng Kuang*
Department of Computer Science
The University of Hong Kong
kuangqipeng@connect.hku.hk

Mengshi ZHAO
Department of Computer Science
The University of Hong Kong
zmsxsl@connect.hku.hk

Abstract

Pedestrian search is a popular fast-growing topic in computer vision. It aims to find the most likely person in a graph to match a given target. In this work, we review some of the popular method, and list some key point of each one. Then we ran the code of a Seqset paper ouselves, to show that how it works and do some further performance analysis.

1 Introduction

Pedestrian search aims to localize and identify pedestrians from a series of uncropped images, incorporating two subtasks of pedestrian detection and pedestrian re-identification. This task was first proposed at the ACM Multimedia Conference in 2013, and for the first time, pedestrian detection and pedestrian re-identification tasks are integrated into one task.

Pedestrian detection is a sub-task of target detection, which aims to find the location and size of a pedestrian from a large amount of photo or video data, and usually requires the use of a rectangular box to frame it out, while this task does not require identifying who the framed pedestrian is; pedestrian re-identification is a task of pedestrian matching, which usually matches a person that most closely resembles the target to be matched from a series of already cropped pedestrian images. The combination of pedestrian detection and pedestrian re-identification makes the pedestrian search task more practical and can be applied in criminal investigations, thus saving a lot of human resources costs.

Since the task of pedestrian search combines two independent subtasks, there are usually two types of solutions: two-stage models and end-to-end models. The two-stage model solves the pedestrian search task in steps: first, the pedestrians in the images are detected using deep learning methods, and for the detected models, the matching task is further completed using a pedestrian re-identification model. The two-stage model requires two separate subtasks, which makes the original task less difficult but also makes the pedestrian search task less efficient, so the end-to-end pedestrian search algorithm has received more attention and research. The end-to-end pedestrian search network centralizes the pedestrian detection and pedestrian re-identification into one network, so that the training and inference of the network can be done for one network, and such a model will be more practical.

2 related work

Taking the common datasets in the field of pedestrian search (CUHK-SYSU, PRW) as an example, the models and methods that perform well on pedestrian search are as follows (data as of November 7, 2022).

*For further information please visit Qipeng's personal webseite kqp.world

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Rank	Model	MAP↑	Top-1	Paper	Code	Result	Year	Tags
1	SeqNeXt+GFN	96.4	97.0	Gallery Filter Network for Person Search			2022	
2	SeqNeXt	96.1	96.5	Gallery Filter Network for Person Search			2022	
3	GLCNet+CBGM	95.8	96.2	Global-Local Context Network for Person Search			2021	
4	GLCNet	95.5	96.1	Global-Local Context Network for Person Search			2021	
5	ROI-AlignPS	95.4	96.0	Efficient Person Search: An Anchor-Free Approach			2021	
6	NAE+SeqNet+CBGM	94.8	95.7	Sequential End-to-end Network for Efficient Person Search			2021	
7	OIM+SeqNet+CBGM	94.3	95.0	Sequential End-to-end Network for Efficient Person Search			2021	
8	AlignPS+	94	94.5	Anchor-Free Person Search			2021	
9	NAE+SeqNet	93.8	94.6	Sequential End-to-end Network for Efficient Person Search			2021	
10	OIM+SeqNet	93.4	94.1	Sequential End-to-end Network for Efficient Person Search			2021	
11	AlignPS	93.1	93.4	Anchor-Free Person Search			2021	

Figure 1: CUHK-SYSU

Rank	Model	mAP↑	Top-1	Paper	Code	Result	Year
1	SeqNeXt+GFN	58.3	92.4	Gallery Filter Network for Person Search			2022
2	SeqNeXt	57.6	89.5	Gallery Filter Network for Person Search			2022
3	ROI-AlignPS	51.6	84.4	Efficient Person Search: An Anchor-Free Approach			2021
4	GLCNet+CBGM	47.8	87.8	Global-Local Context Network for Person Search			2021
5	NAE+SeqNet+CBGM	47.6	87.6	Sequential End-to-end Network for Efficient Person Search			2021
6	GLCNet	46.7	84.9	Global-Local Context Network for Person Search			2021
7	NAE+SeqNet	46.7	83.4	Sequential End-to-end Network for Efficient Person Search			2021
8	OIM+SeqNet+CBGM	46.6	84.9	Sequential End-to-end Network for Efficient Person Search			2021
9	AlignPS+	46.1	82.1	Anchor-Free Person Search			2021
10	AlignPS	45.9	81.9	Anchor-Free Person Search			2021
11	OIM+SeqNet	45.8	81.7	Sequential End-to-end Network for Efficient Person Search			2021

Figure 2: PRW

From the benchmark of both datasets, we can see that all the models/methods are based on SeqNet (including SeqNeXt and GLCNet, which are at the top of the ranking) except for the Anchor-free one. For these two reasons, we intend to reproduce the work of SeqNet.

3 Algorithm introduction

3.1 RCNN

RCNN is a classic approach in the field of target detection, from the 2014 paper "Rich feature hierarchies for accurate object detection and semantic segmentation". RCNN divides target detection into two sub-tasks that are performed in parallel, classification task refers to classifying the content in the bounding box, and regression task refers to finding the exact location of the bounding box by the idea of regression. The inference process consists of the following four steps.

1. generate area proposals (Proposals) using the Selective Search algorithm.
2. scaling down the proposals with different area sizes to a uniform size.
3. inputting the obtained results into a feature extraction network to obtain feature vectors.

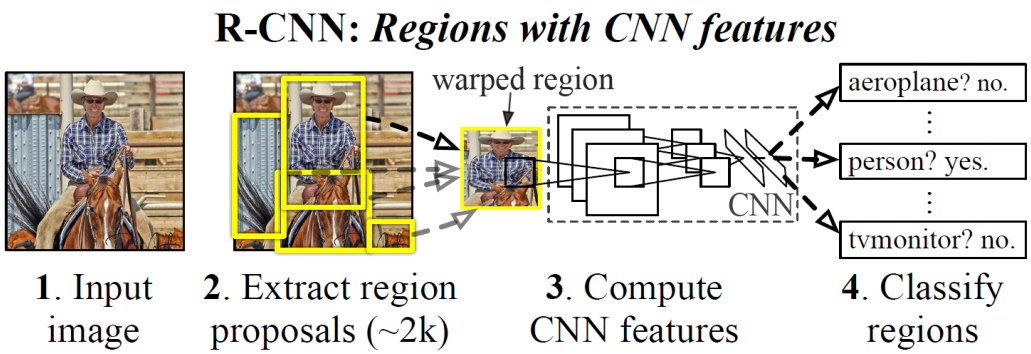


Figure 3: RCNN

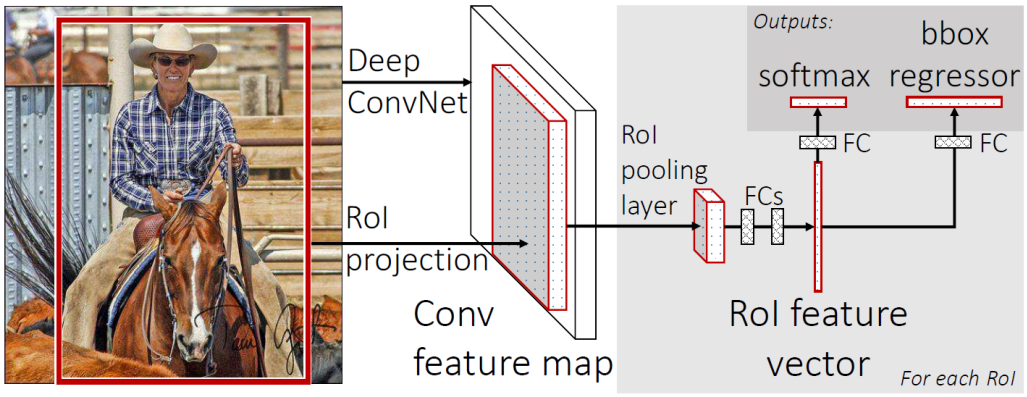


Figure 4: FRCNN

4. use the classifier to get the class of the area and the regression value of bboxes.

The schematic diagram of the algorithm is Figure.3. In practice, for a single image, the RCNN approach has to independently input the extracted two thousand regions into the feature extraction network independently and requires training a complex SVM classifier, so the efficiency of the method is low.

3.2 FastRCNN

Fast RCNN is an improvement of RCNN, from the 2015 paper Fast R-CNN. Fast RCNN proposes the method of ROI pooling, which makes it unnecessary to input the Proposals obtained by Selective Search into the feature extraction network separately, and secondly, it is possible to process these features in parallel. The inference process of the method consists of the following steps.

1. generate region proposals (Proposals) using the Selective Search algorithm.
2. inputting the original image into the feature extraction network to obtain a feature map (feature map) of the whole image, corresponding the proposals to the regions on the feature map, and pooling them to the specified size.
3. extracting the obtained results as ROI feature vectors using a two-layer fully connected network.
4. using softmax and bbox regressor to obtain the results.

The schematic diagram of the algorithm is Figure.4.

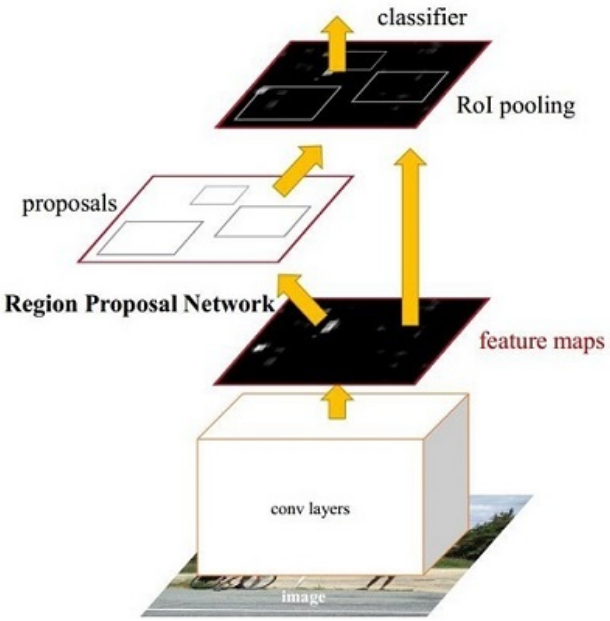


Figure 5: FasterRCNN

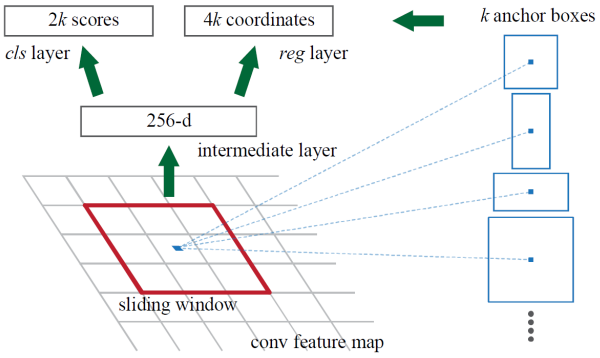


Figure 6: RPN

3.3 FasterRCNN

Faster RCNN by is a further improvement of Fast-RCNN from the 2016 paper "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks".

The outstanding contribution of this method is the proposed RPN (region proposal network) network, which makes it no longer necessary to generate region proposals using the Selective Search algorithm.The network structure of Faster RCNN is composed of Fast RCNN and RPN is Figure.5, where the structure of RPN is Figure.6.

The inference process is as follows: first, on the feature map, a 256-dimensional feature vector is extracted using a 3×3 sliding window, and the classification and regression results of a set of anchor boxes are obtained using this feature vector. A set of anchor boxes contains 9 anchors, corresponding to nine regions of different sizes on the original map 1:1 1:2 2:1×1282 2562 5122, as shown in the Figure.7.

The regions that exceed the original image boundaries are then excluded, and their foreground/background categories and regression parameters are computed to obtain the region proposal proposals, and the highly similar proposals are then excluded using a non-maximal suppression algorithm (NMS), and the final remaining proposals are fed into the original Fast RCNN.

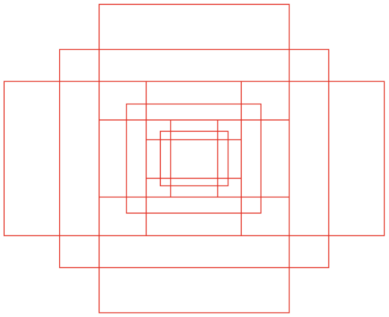


Figure 7: Feature map

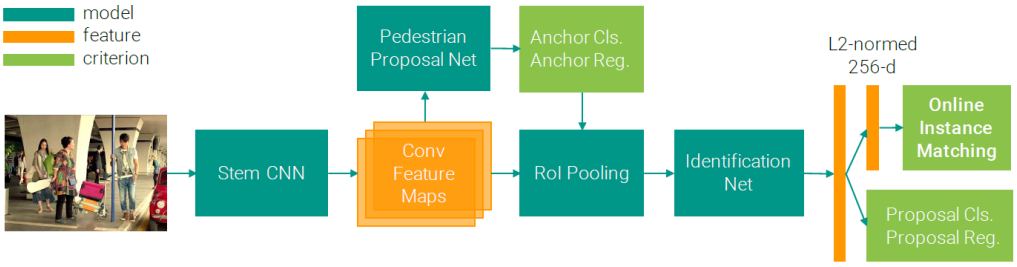


Figure 8: OIM

The RPN structure of the Faster RCNN allows the model to generate more accurate region proposals by itself, and the RPN structure can be trained jointly with the Fast RCNN, which makes the target detection truly an end-to-end algorithm. Meanwhile, in Faster RCNN, the feature extraction network uses a VGG network with better performance, which makes the accuracy of target detection improved.

3.4 ROI-Align

ROI-Align is a modification of ROI pooling, which comes from the 2017 paper "Mask R-CNN".

Since the input of ROI pooling is the coordinates of proposals, which are computed by RPN, and since after one regression on anchor boxes, the proposal maps to the feature map as a floating-point coordinate, which often cannot correspond to a specific pixel, a rounding is needed to let proposals are mapped to the feature map. When calculating ROI pooling, another rounding is needed to map it to the specified size. These two rounding operations actually make the original area corresponding to the features obtained by ROI pooling differ significantly from the proposals, resulting in poor performance of target detection.

ROI-Align uses a bilinear interpolation method to calculate the value of each floating point coordinate on the feature map, thus performing similar ROI pooling without large errors.

3.5 OIM

OIM is a loss of pedestrian search network, which is from the paper Joint Detection and Identification Feature Learning for Person Search.

Its corresponding network structure is shown in the Figure.8.

Compared with Faster RCNN, the difference is that the ROI feature vector originally extracted by this network is downscaled to 256 dimensions from another branch using a fully connected layer, and L2 normalization is performed, and this result is used for Online Instance Matching (OIM). In pedestrian search, the portraits in the query need to be matched with the detected people in the gallery, and the portraits in the query usually have many categories, so using softmax is not very suitable. In model inference, the model first saves the feature vectors extracted from the images in query to the LUT

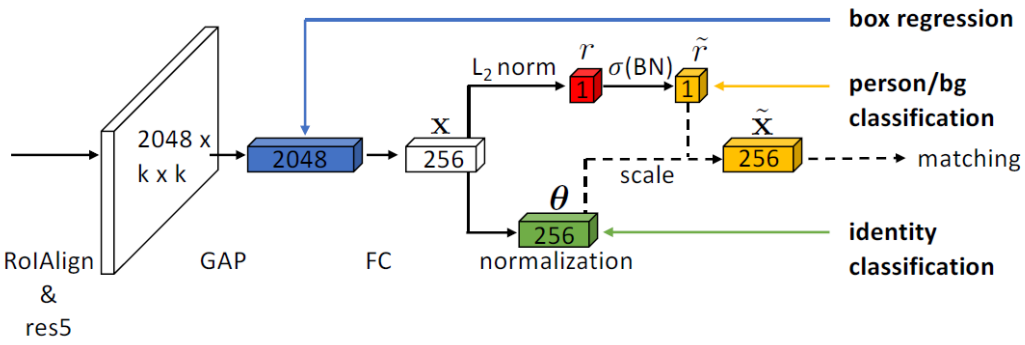


Figure 9: NAE

(lookup table) through the ReID phase of the network, and then passes the images in gallery through the entire network, calculates the feature vectors of the people in the proposal area, and compares the cosine similarity with the ones in the LUT one by one to get the most matching people. The feature vectors of the most matching records in the LUT are updated online at the same time, and perform L2 normalization:

$$\hat{v}_t = \gamma v_t + (1-\gamma)x \tag{1}$$

In the model training, the loss using OIM comes from two parts, the idea is to close the distance of the feature vectors of the same pedestrian and farther the distance of the feature vectors between different pedestrians, first calculate the cosine similarity between the pedestrian in the framed and the labeled foreground, then calculate the cosine similarity between the pedestrian in the framed and the unlabeled foreground, and construct the softmax probability:

$$\begin{aligned} p_i &= \frac{\exp\left(v_i^T x/\tau\right)}{\sum_{j=1}^{L}\exp\left(v_j^T x/\tau\right)+\sum_{k=1}^{Q}\exp\left(u_k^T x/\tau\right)} \\ q_i &= \frac{\exp\left(u_i^T x/\tau\right)}{\sum_{j=1}^{L}\exp\left(v_j^T x/\tau\right)+\sum_{k=1}^{Q}\exp\left(u_k^T x/\tau\right)} \end{aligned} \tag{2}$$

where the LUT still needs to be updated online, and the unlabeled foreground is the one in the most recent batch. The loss function for the OIM part is then constructed as $\mathcal{L} = E_x[logp_t]$. And the entire network is trained jointly under the supervision of four losses.

3.6 NAE

NAE also corresponds to the loss of a specific pedestrian search network from the 20-year paper "Norm-Aware Embedding for Efficient Person Search". The structure of the network corresponding to this method is illustrated in Figure.9. The difference between NAE and OIM network is the features used for foreground/background classification. OIM uses the ROI feature vector, which does not reflect the foreground/background information well, so based on OIM, NAE network uses the length of the extracted L_2 normalized feature vector as the basis for foreground/background classification, and this metric can reflect the foreground/background information well.

The formula for L2 normalization of the feature vector x is $x = r \cdot \theta$, where θ is the unit length vector after L_2 normalization, which is consistent with OIM, and this vector is used for the training process to match the target portrait. And r as the length of x, which can represent the foreground/background information well after the BN operation, due to the matching similarity is calculated as $\text{sim}\left(\tilde{\mathbf{x}}_q, \tilde{\mathbf{x}}_g\right) = \tilde{\mathbf{x}}_q^T \tilde{\mathbf{x}}_g = \tilde{r}_g \cdot \boldsymbol{\theta}_q^T \boldsymbol{\theta}_g$.

The closer r is to 1, representing the foreground, the higher the matching similarity at that time; the closer r is to 0, representing the background, the lower the matching similarity at that time. For the training of r cross entropy is used as supervision:

$$\mathcal{L}_{\text{det}} = -y\log(\tilde{r}) - (1-y)\log(1-\tilde{r})$$

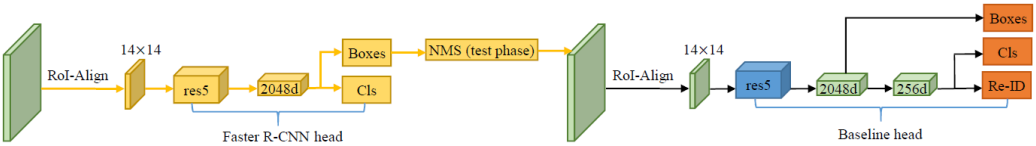


Figure 10: SeqNet

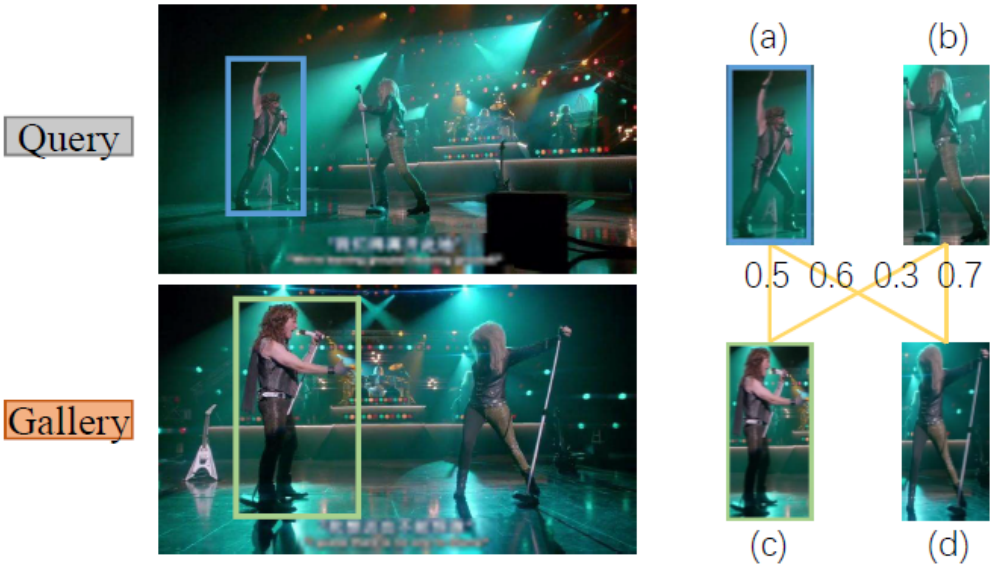


Figure 11: SeqNet-example

The NAE network further elaborates the foreground/background classification supervision based on OIM, making the cosine similarity of feature vector x more meaningful during the matching process. Thus, the accuracy of pedestrian search is further improved.

3.7 SeqNet

SeqNet is the serialized pedestrian search network from the paper "Sequential End-to-end Network for Efficient Person Search" in ’21, which became the SOTA at that time with 94.8 mAP on CUHK-SYSU dataset. SeqNet adds two improvements to NAE, one is growing the network structure, which allows the model to use more accurate proposals for subsequent ReIDs, as shown in Figure.10.

It can be seen that the proposals used in the ReID stage are more accurate proposals output by the full Faster RCNN regression, which undoubtedly provides higher quality region proposals for pedestrian matching, which allows the end-to-end model accuracy to catch up with the accuracy of the two-stage model. Another improvement is that SeqNet proposes a contextual bipartite graph matching (CBGM) method, which makes full use of the contextual information in the image, i.e., the information of other pedestrians that may be contained in the image, so that the matching process is not performed singularly with the boxed out people against all the people to be matched, but considers the best match of all people in the whole image against the people to be matched.

If only the matching of (a) with (c) and (d) is considered, then (a) will be incorrectly matched with a higher probability of (d); however, if the matching of the whole image of Query is considered, i.e., if the matching of (a) and (b) with (c) and (d) is considered, then (b) and (d) will be matched first, at which point (a) will be correctly matched with (c). This problem can be abstracted as the best matching problem for bipartite graphs and solved using the classical Kuhn-Munkres algorithm.

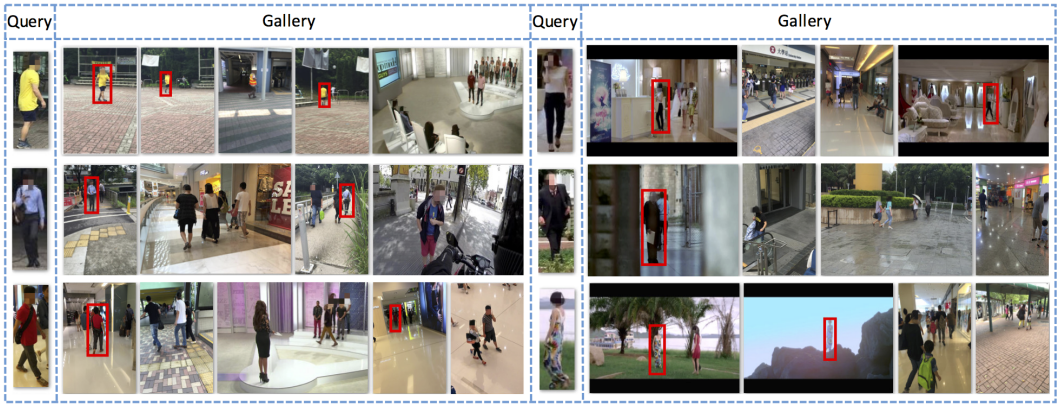


Figure 12: CUHK-dataset

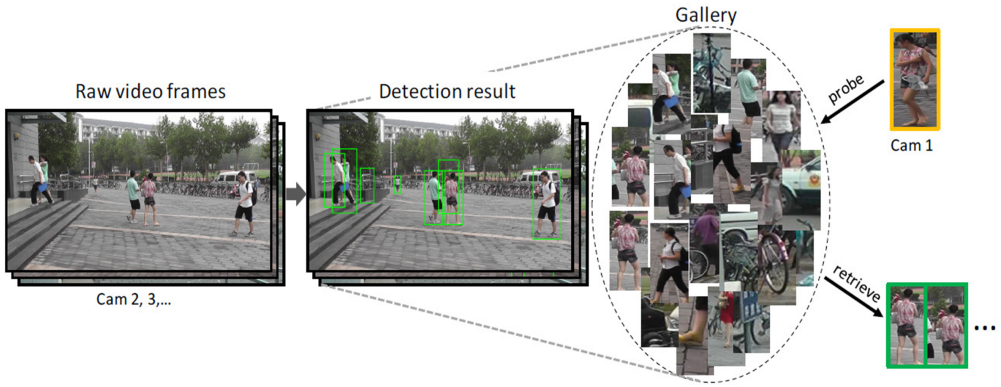


Figure 13: PRW-dataset

4 Experiment

4.1 Dataset

CUHK-SYSU is the most commonly used dataset in the field of pedestrian search.

The dataset is a large-scale benchmark for people search, containing 18,184 images and 8,432 identities. Depending on the image source, the dataset can be divided into two parts: street captures and movies. In street capture, images are collected via handheld cameras across hundreds of scenes and attempt to include variations in point of view, lighting, resolution, and occlusion; in addition, the dataset selects film and TV shows as another source of image capture because they provide more diverse scenes and more challenging perspectives.

The dataset provides annotations for person re-identification and pedestrian detection. Each query person appears in at least two images, and each image can contain multiple query persons and multiple background persons. The data is divided into a training set and a test set. The training set contains 11206 images and 5532 query persons, and the test set contains 6978 images and 2900 query persons. Examples of the dataset is in Figure.12.

PRW is another common dataset for pedestrian search.

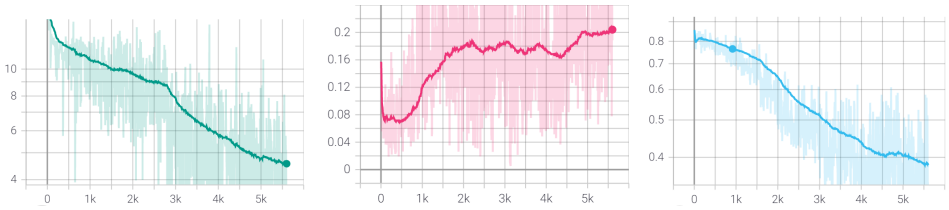
PRW (Person Re-identification in the Wild) is a person re-identification dataset. The dataset was collected at Tsinghua University, and a total of 10 hours of video was captured through six cameras. The dataset is divided into training, validation and test sets. The training set contains 5134 frames and 482 IDs, the validation set contains 570 frames and 482 IDs, and the test set contains 6112 frames and 450 IDs. all pedestrians appearing in each frame are labeled with a bounding box and assigned an ID. Examples of the dataset is in Figure.13.

4.2 Parameters

See Table.1.

4.3 Loss function

See Figure.14.



(a) REID loss (b) Box Classified losses (c) Box Regression losses

Figure 14: Loss function

4.4 Experiment result

4.5 Instance

To see how the algorithm work, Figure. We give a example here. The query is in Figure.15, and the results are in Figure.16.

References

Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12615–12624, 2020.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Figure 15: Query

Table 1: Parameter table

Parameter	Value
Input of Model	
C.INPUT.DATASET	"CUHK-SYSU"
C.INPUT.DATA_ROOT	"data/CUHK-SYSU"
C.INPUT.MIN_SIZE	900
C.INPUT.MAX_SIZE	1500
C.INPUT.BATCH_SIZE_TRAIN	5
C.INPUT.BATCH_SIZE_TEST	1
C.INPUT.NUM_WORKERS_TRAIN	5
C.INPUT.NUM_WORKERS_TEST	1
Learning Parameters	
C.SOLVER.MAX_EPOCHS	20
C.SOLVER.BASE_LR	0.003
C.SOLVER.WARMUP_FACTOR	1.0/1000
C.SOLVER.WEIGHT_DECAY	0.0005
C.SOLVER.SGD_MOMENTUM	0.9
Loss function parameters	
C.SOLVER.LW_RPN_REG	1
C.SOLVER.LW_RPN_CLS	1
C.SOLVER.LW_PROPOSAL_REG	10
C.SOLVER.LW_PROPOSAL_CLS	1
C.SOLVER.LW_BOX_REG	1
C.SOLVER.LW_BOX_CLS	1
C.SOLVER.LW_BOX_REID	1
RPN parameters	
C.MODEL.RPN.NMS_THRESH	0.7
C.MODEL.RPN.BATCH_SIZE_TRAIN	256
C.MODEL.RPN.POS_FRAC_TRAIN	0.5
C.MODEL.RPN.POS_THRESH_TRAIN	0.7
C.MODEL.RPN.NEG_THRESH_TRAIN	0.3
C.MODEL.RPN.PRE_NMS_TOPN_TRAIN	12000
C.MODEL.RPN.PRE_NMS_TOPN_TEST	6000
C.MODEL.RPN.POST_NMS_TOPN_TRAIN	2000
C.MODEL.RPN.POST_NMS_TOPN_TEST	300
ROI parameters	
C.MODEL.ROI_HEAD.BN_NECK	TRUE
C.MODEL.ROI_HEAD.BATCH_SIZE_TRAIN	128
C.MODEL.ROI_HEAD.POS_FRAC_TRAIN	0.5
C.MODEL.ROI_HEAD.POS_THRESH_TRAIN	0.5
C.MODEL.ROI_HEAD.NEG_THRESH_TRAIN	0.5
OIM parameters	
C.MODEL.LOSS.LUT_SIZE	5532
C.MODEL.LOSS.CQ_SIZE	5000
C.MODEL.LOSS.OIM_MOMENTUM	0.5

Table 2: Experiment result

epoch	recall	AP	mAP	Top-1	Top-5	Top-10
1	34.65%	26.60%	41.74%	45.31%	59.10%	62.93%
2	63.73%	60.36%	79.05%	81.10%	91.38%	93.31%
3	84.19%	79.23%	87.69%	88.59%	96.52%	97.69%
4	87.37%	82.37%	90.02%	91.45%	96.93%	97.86%
5	88.66%	83.79%	90.49%	91.55%	97.28%	98.21%
6	89.73%	84.59%	91.10%	91.76%	97.21%	98.38%

Figure 16: Gallery

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Zhengjia Li and Duoqian Miao. Sequential end-to-end network for efficient person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2011–2019, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

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