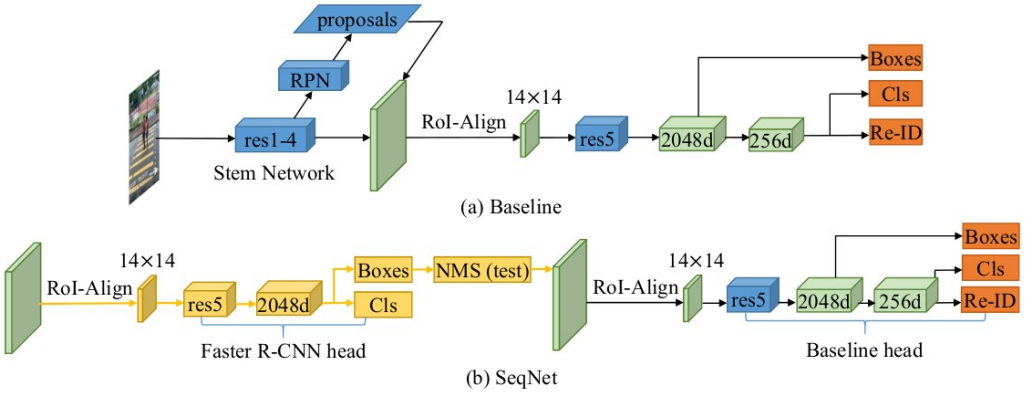
First Reading Pass - 17.10.21

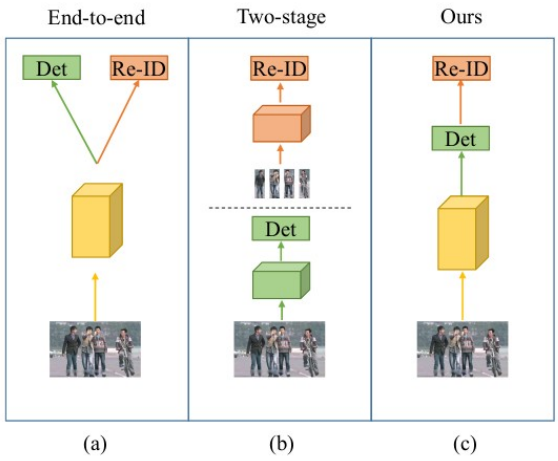
Li, Z., & Miao, D., 2021, Sequential End-to-end Network for Efficient Person Search, AAAI.

Summary

This article aims to improve existing methods on person search. Parallel end to end networks have already been found using similar technologies as mentioned in this article, but they have given low quality bounding boxes. Therefore, in this article, it is tried to make a more efficient person search by using a sequential end-to-end network. The network baseline consists of two basic elements. The first is pedestrian detection and the second is re-ID person identification. There is a brief summary of the model in the figure below.



The comparison with other similar studies can be seen below figure.



Two different datasets were used in this study, CUHK-SYSU (Xiao et al. 2017) and PRW (Zheng et al. 2017). Details are available in the data properties section. In the evaluation part, the performance measures are the Cumulative Matching Characteristic (CMC) and the average Average Precision (mAP). It is aimed to maximize both metrics.

Glossary

Region Proposal Network

A Region Proposal Network, or RPN, is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end to generate high-quality region proposals. RPN and algorithms like Fast R-CNN can be merged into a single network by sharing their convolutional features - using the recently popular terminology of neural networks with attention mechanisms, the RPN component tells the unified network where to look.

Non-Maximum Suppression

Non-Maximum Suppression (NMS) is a technique used in numerous computer vision tasks. It is a class of algorithms to select one entity (e.g., bounding boxes) out of many overlapping entities. We can choose the selection criteria to arrive at the desired results.

Cross-Level Semantic Alignment

The CLSA contains two components:

1- Person detection which locates all person instances in the gallery scene images for facilitating the subsequent identity matching.

2- Person re-identification which matches the probe image against a large number of arbitrary scale gallery person bounding boxes (the key component of CLSA).

** There are more keywords in this article. However, these keywords are the main topic of reference articles. That’s why, I am not sure if they are suitable for glossary and I leave it blank.

Past Work

Lan, X., Zhu, X., & Gong, S. (2018). Person Search by Multi-Scale Matching. ECCV.

Wang, C., Ma, B., Chang, H., Shan, S., & Chen, X. (2020). TCTS: A Task-Consistent Two-Stage Framework for Person Search. CVPR, 11949-11958.

Dong, W., Zhang, Z., Song, C., & Tan, T. (2020). Instance Guided Proposal Network for

Person Search. CVPR, 2582-2591.

Related Work

Han, B., Ko, K., & Sim, J. (2021). Context-Aware Unsupervised Clustering for Person Search. ArXiv, abs/2110.01341.

Yan, Y., Li, J., Qin, J., Liao, S., & Yang, X. (2021). Efficient Person Search: An Anchor-Free Approach. ArXiv, abs/2109.00211.

Yan, Y., Li, J., Liao, S., Qin, J., Ni, B., Yang, X., & Shao, L. (2021). Exploring Visual Context for Weakly Supervised Person Search. ArXiv, abs/2106.10506.

Data Properties

Two different datasets are used in this article. The details of these two articles are as follows.

CUHK-SYSU

CUHK-SYSU (Xiao et al. 2017) is a large scale person search dataset containing 18,184 scene images and 96,143 annotated BBoxes, which are collected from two sources: street snap and movie. All people are divided into 8,432 labeled identities and other unknown ones. The training set contains 11,206 images and 5,532 different identities. The test set contains 6,978 images and 2,900 query people. The training and test sets have no overlap on images and query people. For each query, different gallery sizes from 50 to 4000 are pre-defined to evaluate the search performance. If not specified, the gallery size of 100 is used by default.

PRW

PRW is another widely used dataset (Zheng et al. 2017) containing 11,816 video frames captured by 6 cameras in Tsinghua university. 34,304 BBoxes are annotated manually. Similar to CUHK-SYSU, all people are divided into labeled and unlabeled identities. The training set contains 5,704 images and 482 different people, while the test set includes 6,112 images and 2,057 query people. For each query, the gallery is the whole test set, i.e., the gallery size is 6112.