END-TO-END PERSON SEARCH SEQUENTIALLY TRAINED ON AGGREGATED DATASET

Angelique Loesch, Jaonary Rabarisoa, Romaric Audigier
CEA LIST, Vision and Learning Lab for Scene Analysis, France / Vision Lab, Thales SIX GTS, France
{angelique.loesch, jaonary.rabarisoa, romaric.audigier}@cea.fr

MOTIVATIONS

- In **real use-case scenarios**,
  - Extraction of human snippets from full scene images
  - **Re-ID depending** on the quality of a person **detector**
- **Person search** : problem considering **both detection and re-ID** tasks in a **unique framework**
  - Training dataset with annotated bounding boxes and IDs
  - **Difficult** to collect datasets **with both annotation types**

CONTRIBUTIONS

- A **new end-to-end CNN model** reaching state-of-the-art accuracy
- A **study on the tradeoff between runtime and performance** w.r.t. the shared backbone size
- A **sequential training** with **aggregation of more train datasets** for **people detection ➔ Improvement of re-ID performance**
  - in **intra-dataset** scenarios
  - in **cross-dataset** scenarios, of utmost importance for real use-cases

PERSON SEARCH DATASETS

- PRW dataset [4]
  - 11,8k images with 43,1k boxes (8,8k distractors)
  - 932 IDs

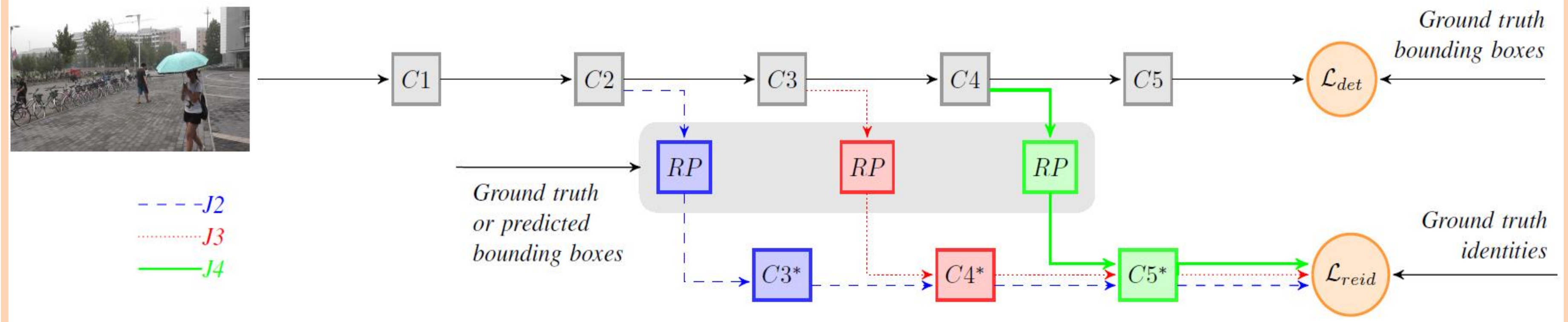
- CUHK-SYSU dataset [21]
  - 18,1k images with 99,8k boxes
  - 8,4k IDs

[4] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," in IEEE CVPR, 2017.
[21] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in IEEE CVPR, 2016.

KEYWORDS

Person Detection   Person Search   Re-identification
Multi-Task Learning   Cross-Dataset

PROPOSED METHOD



- An **SSD** architecture keeping the **performance** of the **detection task as high as possible**
- A **maximum number of shared layers** between the detection and the re-ID branches **to reduce forward complexity**.
- **A triplet loss** to solve the re-ID task as it is an **effective way to learn representation**
- A **two-step sequential training** to **exploit all available detection data along with joint** detection and re-ID annotated **data**:
  - Training detection branch only
  - Training re-ID branch by freezing the common layers

RESULTS

❖ **Comparison with person search state-of-the-art**

➢ On CUHK-SYSU, **top-2** or **top-3** best mAP

➢ On PRW, **top-1** best mAP

	PRW		CUHK-SYSU gallery size 100 / 4000	
	mAP (%)	Rank-1 (%)	mAP (%)	Rank-1 (%)
*J*2 (ours)‡	**25.2**	47.0	76.4 / 49.2	76.7 / 51.3
*J*3 (ours)‡	22.5	45.1	79.4 / 55.8	80.5 / **58.9**
*J*4 (ours)‡	12.3	27.3	76.7 / 53.3	77.8 / 56.0
Xiao2016 [14]	-	-	55.7 / -	62.7 / 42.5
JDI+OIM [15]‡	21.3	49.9	75.5 / 51.0	78.7 / -
IAN [16]*	23.0	61.8	77.2 / 55.0	80.7 / -
Chen 2018 [17]	-	-	78.8 / -	80.9 / -
I-Net [18]	-	-	79.5 / 53.5	**81.5** / -
Liu2018 [19]*	21.0	63.1	**79.8** / -	79.9 / -
JDI+IEL [20]*	24.3	**69.5**	79.4 / **58.0**	79.7 / -
NPSM [11]‡	24.2	53.1	77.9 / 54.0	81.2 / -

Highest score reported for PRW protocol at
*: 3 bounding boxes / image; ‡: 5 bounding boxes / image.

Mean average precision (mAP) (%) and matching rate at rank-1 (Rank-1) (%)

[18] Z. He, L. Zhang, and W. Jia, "End-to-end detection and re-identification integrated net for person search," in arXiv preprint arXiv:1804.00376, 2018.
[19] H. Liu, W. Shi, W. Huang, and Q. Guan, "A discriminatively learned feature embedding based on multi-loss fusion for person search," in IEEE ICASSP, 2018.
[20] W. Shi, H. Liu, F. Meng, and W. Huang, "Instance enhancing loss: Deep identity-sensitive feature embedding for person search," in IEEE ICIP, 2018.

EXPERIMENTS

❖ **Influence of shared backbone size on re-ID performance**

	PRW		CUHK-SYSU gallery size 100 / 4000	
	mAP (%)	Rank-1 (%)	mAP (%)	Rank-1 (%)
Disj.‡	13.3	32.3	72.1 / 50.1	74.1 / 53.3
*J*2 (ours)‡	**25.2**	47.0	76.4 / 49.2	76.7 / 51.3
*J*3 (ours)‡	22.5	45.1	79.4 / 55.8	80.5 / **58.9**
*J*4 (ours)‡	12.3	27.3	76.7 / 53.3	77.8 / 56.0

Highest score reported for PRW protocol at
*: 3 bounding boxes / image; ‡: 5 bounding boxes / image.

mAP (%) and Rank-1 (%) on PRW and CUHK-SYSU

➢ Best trade-off accuracy/ running time on both datasets for medium-sized backbone.

❖ **Influence of shared backbone size on computation time**

	#im. batch	computation time (ms) 5 p. / im.	computation time (ms) 20 p. / im.
Disj.	1	17.0	7.4
	4	13.6	6.6
	8	12.9	6.4
*J*2	1	12.3	3.9
	4	8.3	2.7
	8	7.4	2.5
*J*3	1	12.2	3.4
	4	7.8	2.4
	8	7.2	2.2
*J*4	1	**11.1**	**2.9**
	4	**7.1**	**1.9**
	8	**6.5**	**1.9**

Mean computation time (ms) to detect a person and extract his/her feature

➢ Up to **3.4 faster** than disjoint architecture

Boosting shared feature map efficiency

❖ **for intra-dataset scenarios**

	gallery size 100 / 4000			
	mAP (%)	Rank-1 (%)	mAP GT (%)	Rank-1 GT (%)
$J2_c$	71.4 / 43.6	71.6 / 45.5	78.6 / 50.3	78.0 / 52.3
$J2$	76.4 / 49.2	76.7 / 51.3	81.9 / 54.9	81.0 / 56.5
$J3_c$	75.5 / 48.1	76.4 / 50.3	81.2 / 54.2	80.9 / 56.5
$J3$	**79.4 / 55.8**	**80.5 / 58.9**	**84.4 / 60.9**	**84.0 / 63.1**
$J4_c$	62.9 / 33.3	62.3 / 33.8	68.5 / 37.1	67.1 / 37.2
$J4$	76.7 / 53.3	77.8 / 56.0	81.6 / 57.1	81.3 / 58.8

*(**left**) mAP and Rank-1 on CUHK-SYSU for our joint models trained on CUHK-SYSU only, or boosted by pedestrian dataset aggregation. (**right**) Ground truth boxes instead of predicted boxes*

➢ Greater improvement for longer backbone

➢ Up to +20 p.p. mAP

❖ **for cross-dataset scenarios**

	gallery size 100 / 4000	
	mAP (%)	Rank-1 (%)
$J2_{p}$	31.5 / 14.9	33.4 / 16.1
$J2_{m\text{-}w\text{-}p}$	**54.4 / 29.4**	**55.4 / 31.9**
$J3_{p}$	29.8 / 13.9	31.7 / 15.8
$J3_{m\text{-}w\text{-}p}$	54.6 / 28.1	56.1 / 29.7
$J4_{p}$	22.8 / 8.8	23.1 / 9.5
$J4_{m\text{-}w\text{-}p}$	52.5 / 27.8	53.3 / 28.8

mAP and Rank-1 on CUHK-SYSU cross-dataset for our joint models trained on PRW only, or boosted by pedestrian dataset aggregation

➢ A **not costly yet efficient** way to increase re-ID performance

➢ > 20 p.p. mAP improvement

CONCLUSION

❖ **New end-to-end person search networks** based on
- SSD architecture for detection
- Triplet loss to solve re-ID

❖ **Competitive re-ID results** on CUHK-SYSU and PRW datasets

❖ **Aggregating pedestrian datasets** during **training** leads to **significant improvement in intra and cross-dataset Re-ID** scenarios