

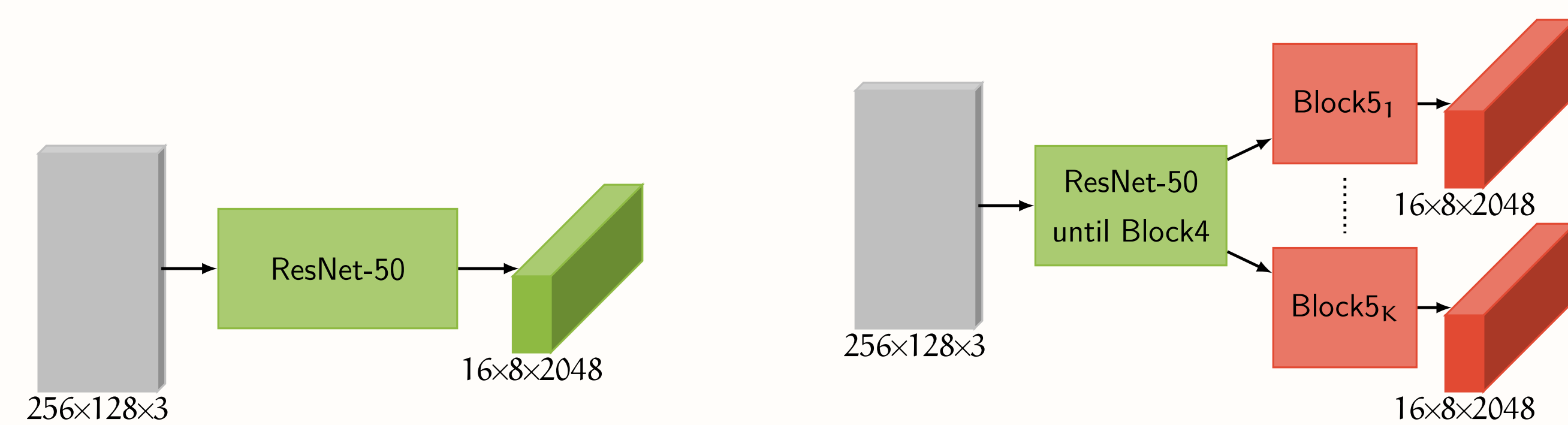


Summary

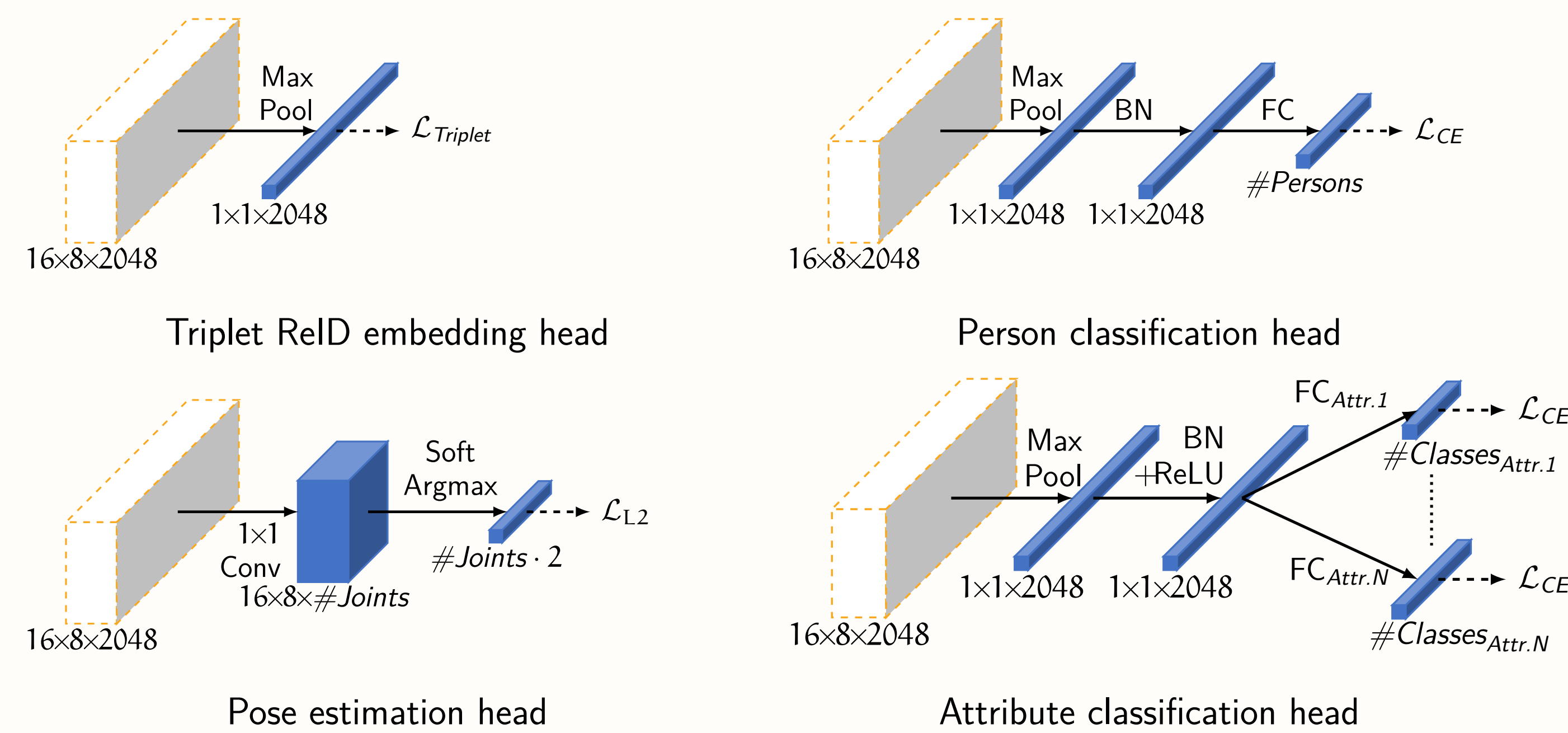
- Mobile vision applications need to perform many person-related tasks. We focus on:
 - Person re-identification (Market-1501)
 - Body part segmentation (LIP)
 - Human pose estimation (MPII, LIP)
 - Attribute classification: gender, clothing etc. (Market-1501)
- These tasks are interdependent and mobile platforms are resource-constrained:
 - Joint multi-task learning is needed.
- There exists no single dataset that provides annotations for all tasks:
 - Option 1: Generate pseudo-labels for a single dataset.
 - Option 2: Combine multiple datasets during training.
- We investigate architectural design choices and their effects on joint training, compared to single-task baselines and the state-of-the-art.

Architecture

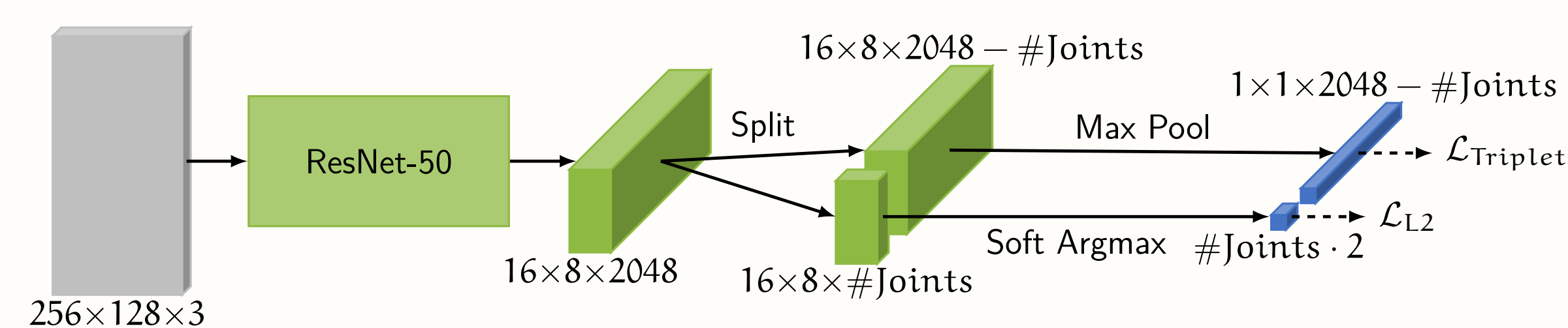
How much to share among tasks?



Individual Task Heads



ReID task ⚡ Pose task



Split ReID and pose head to avoid interference.

Automatic Annotations



Automatic annotations on the Market-1501 dataset for pose estimation (top) and part segmentation (bottom).

Qualitative Results



Given person detections, we can perform all tasks simultaneously with 50 detections/s.

Results

Automatic Annotations

Market		Evaluation				
Tripl.	Clas.	Attr.	Pose	Pose	Segmentation	
			mAP	PCKh	mIoU	mIoU ₅
$\mathcal{M}^{SB}$	✓	✓	78.6	30.8	—	49.6
	✓	✓	79.2	28.8	—	47.9
$\mathcal{M}^{MB}$	✓	✓	77.7	40.4	—	47.9
	✓	✓	78.2	39.7	—	46.7
Baseline			77.4	46.9	—	48.7

Multi-Dataset Learning

Training				Evaluation				
Market	MPII	LIP		Market	MPII	LIP		
Tripl.	Clas.	Attr.	Pose	RelD	Attr.	Pose	Pose	Segmentation
				mAP	acc	PCKh	PCKh	mIoU
$\mathcal{M}^{SB}$	✓		✓	78.0	—	86.8	74.3	49.9
	✓	✓	✓	78.3	87.1	86.7	73.8	49.6
$\mathcal{M}^{MB}$	✓		✓	77.9	—	86.9	75.0	48.5
	✓	✓	✓	— out of GPU memory —				
$\mathcal{M}^{SB/Split}$	✓	✓	✓	79.1	86.7	86.5	74.4	49.6
Baselines				77.4	88.2	86.6	73.9	47.8
SOTA				86.9[5]	89.7[3]	88.5[4]	82.5[2]	54.4[1]

Findings and Conclusions

- GroupNorm > BatchNorm for multi-dataset training.
- Using more training data is beneficial.
- Synergy effects between tasks:
 - ReID & Part Segmentation
 - ReID & Attribute
 - Pose & Part Segmentation

Acknowledgements

This project was funded, in parts, by ERC Consolidator Grant project “Dee-ViSe” (ERC-CoG-2017-773161) and the BMBF projects “FRAME” (16SV7830) and “PARIS” (16ES0602). Istvan Saranci’s research is funded by a grant from the Bosch Research Foundation. Most experiments were performed on the RWTH Aachen University CLAIX 2018 GPU Cluster.

References

- [1] Jin, X., Lan, C., Zeng, W., Zhang, Z., Chen, Z.: CaseNet: Content-Adaptive Scale Interaction Networks for Scene Parsing. arXiv:1904.08170 (2019)
- [2] Liang, X., Gong, K., Shen, X., Lin, L.: Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark. PAMI **41**(4), 871–885 (2018)
- [3] Liu, H., Wu, J., Jiang, J., Qi, M., Bo, R.: Sequence-based Person Attribute Recognition with Joint CTC-Attention Model. W:1811.08115 (2018)
- [4] Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning Feature Pyramids for Human Pose Estimation. In: ICCV (2017)
- [5] Zheng, F., Deng, C., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F., Ji, R.: Pyramidal Person Re-Identification via Multi-Loss Dynamic Training (2019)