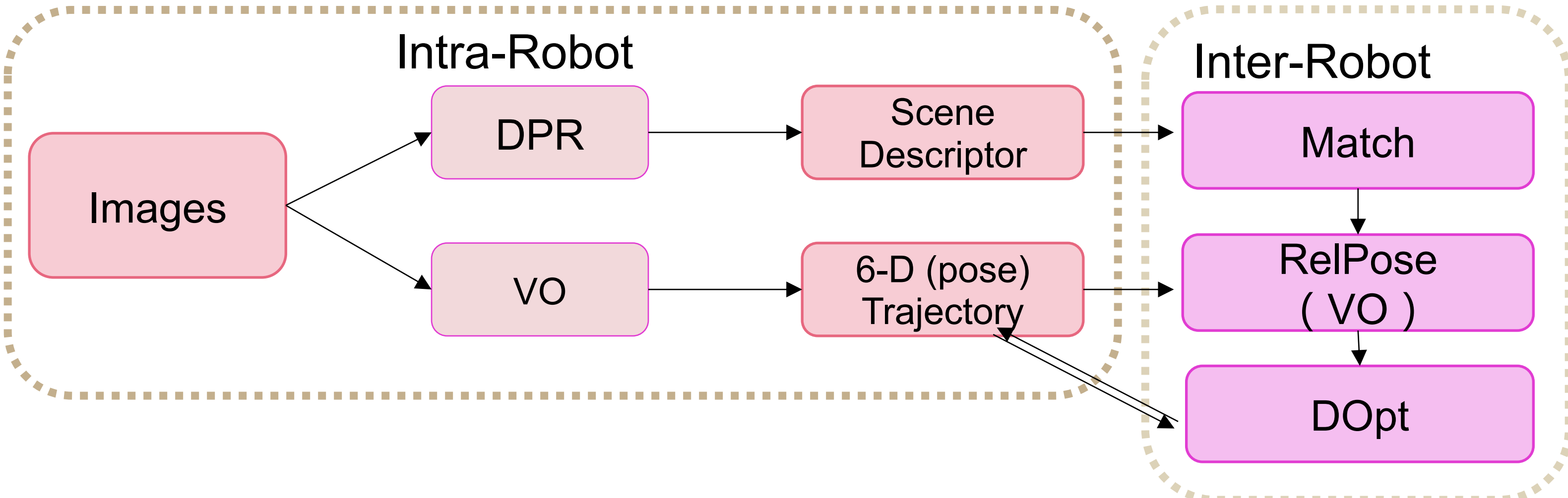
# CNN-based Monocular Decentralized SLAM on Embedded FPGA

Jincheng Yu, Feng Gao, Jianfei Cao, Chao Yu, Zhaoliang Zhang, Zhengfeng Huang, Yu Wang and Huazhong Yang

yjc16@mails.tsinghua.edu.cn
yu-wang@mail.tsinghua.edu.cn

# DSLAM： Basic Task of Multi-robot

- Decentralized visual simultaneous localization and mapping (DSLAM)




  - **DPR:** Decentralized Place Recognition. DPR produces a compact image representation to be communicated among robots.
  - **VO:** Visual Odometry. VO is used to calculate the intra-robot 6-DoF absolute pose from the input frames.
  - **Match, RelPose and DOpt** find out the candidate inter-robot place recognition matches and establish&optimize the relative poses between the matched inter-robot place.

# CNN promotes the DPR and VO

- DPR consumes most of the computation and directly affect the system accuracy.

- **DPR:**
- CNN method greatly **improves the accuracy** of DPR.
  - Traditional BoW+SIFT [1]: 47% match accuracy
  - CNN method NetVLAD[2]: 62% match accuracy

(a) Mobile phone query (b) Retrieved image of same place

CNN can also solve difficult place recognition tasks, such as cross-day-and-night matching [2] .

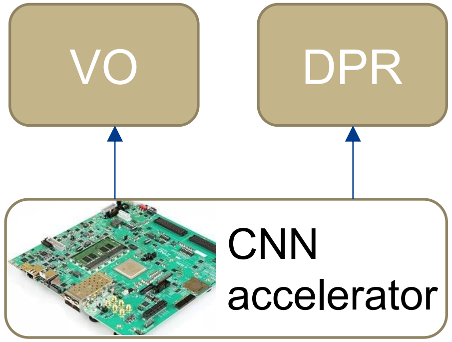

[1]G. Tolias, Y. Avrithis, and H. Juegou, “To aggregate or not to aggregate: Selective match kernels for image search,” in Proc. IEEE Conf. Comput. Vis., 2013, pp. 1401–1408.

[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic,“NetVLAD: CNN architecture for weakly supervised place recognition,”in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 5297–5307.

# CNN promotes the DPR and VO

- DPR and VO consume most of the computation and directly affect system accuracy.
- **VO:**
- Traditional feature point based method:
  - Stereo or RGBD camera: Heavyweight sensors
  - Decrease in accuracy in the absence of texture
  - Design a special accelerator for a method: Lack flexibility.
- CNN method:
  - **[Lightweight]** End-to-end VO from monocular camera
  - **[Robustness]** Applicable in absence of texture
  - **[Flexibility]** VO shares CNN accelerator with other operations


CNN is applicable in the absence of texture[1]

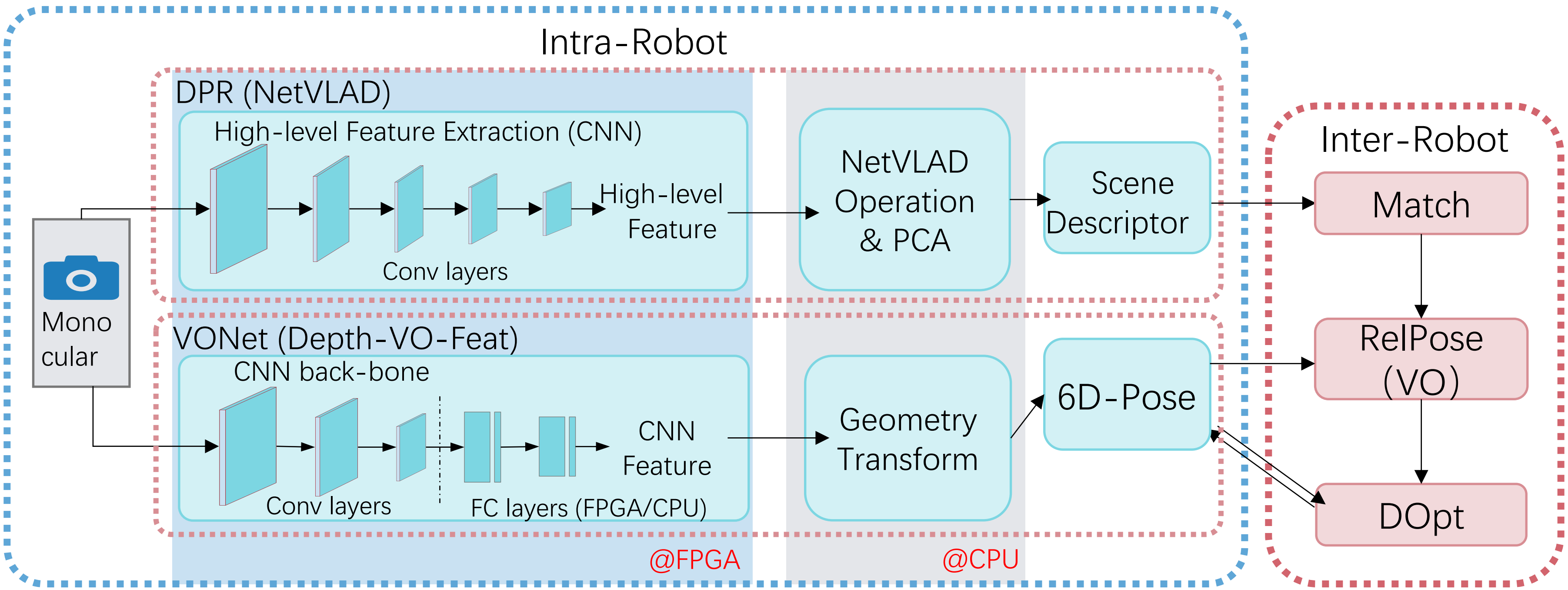



Different CNN methods can be accelerated by a single accelerator

[1] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

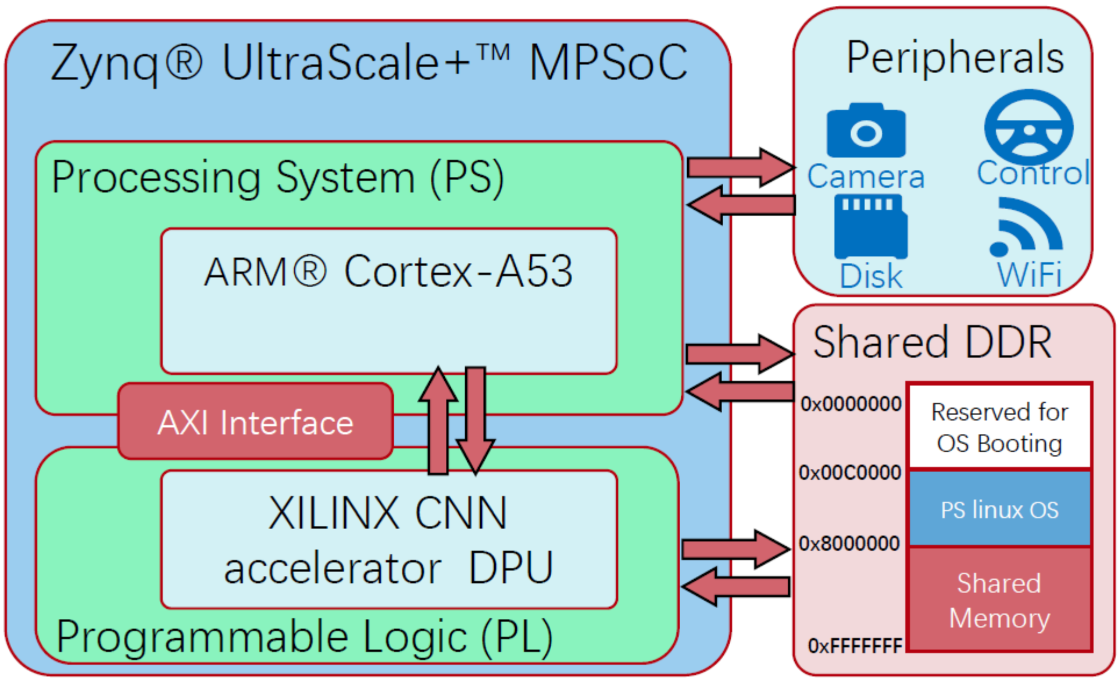
# CNN-based Monocular DSLAM on FPGA

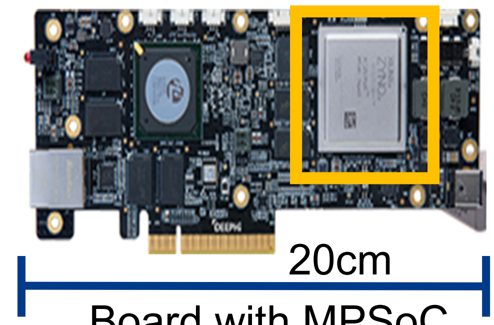
- Implement the computation consuming VO and DPR on Embedded FPGA platform
  - **CNN backbones** of DPR and VO are calculated on CNN accelerator in fixed-point number **at the FPGA side**.
  - **The post-processing operations**, such as PCA and geometry transformation, are calculated **at the CPU side.**



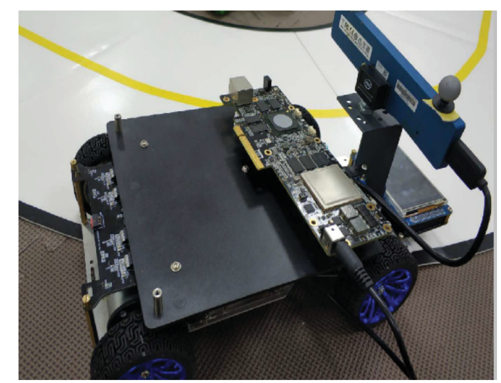
# Embedded FPGA for CNN-based DSLAM

- Use Xilinx UltraScale+ MPSoC [1] to accelerate DSLAM.
  - Perform CNN backbones on the CNN accelerator (DPU[2]) at the FPGA side.
  - Do other operations on the CPU side.




Board with MPSoC

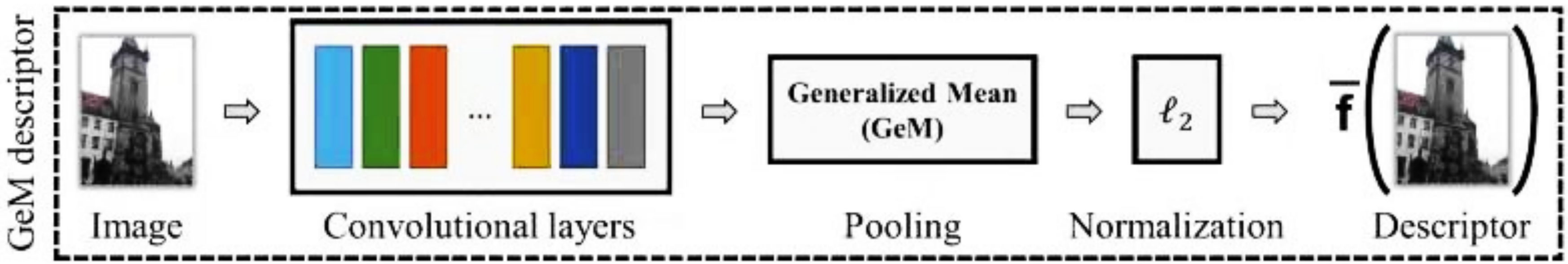


Board on robot

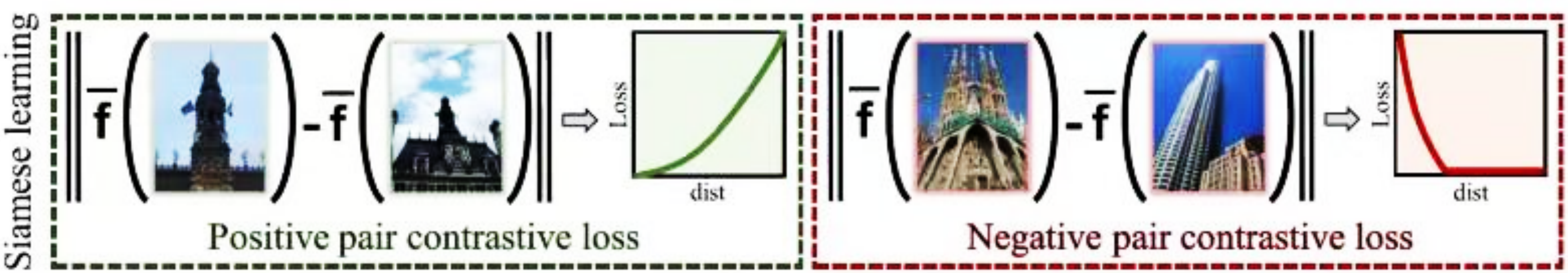

[1] "Xilinx Zynq UltraScale+ MPSoC ZCU102 Evaluation Kit," 2019. [Online]. Available: https://www.xilinx.com/products/boards-and-kits/ ek-u1-zcu102-g.html
[2] "DNNDK User Guide - Xilinx," 2019. [Online]. Available: https://www.xilinx.com/support/documentation/user guides/ug1327-dnndk-user-guide.pdf

# CNN-based Place Recognition

- **Network Structure**
  - CNN backbone and Post-processing (Pooling and Normalization) [1]




- **Data Set** and **Triplet-Loss**[2]
  - The descriptor is more similar, the loss is lower in positive pair and is higher in negative pare

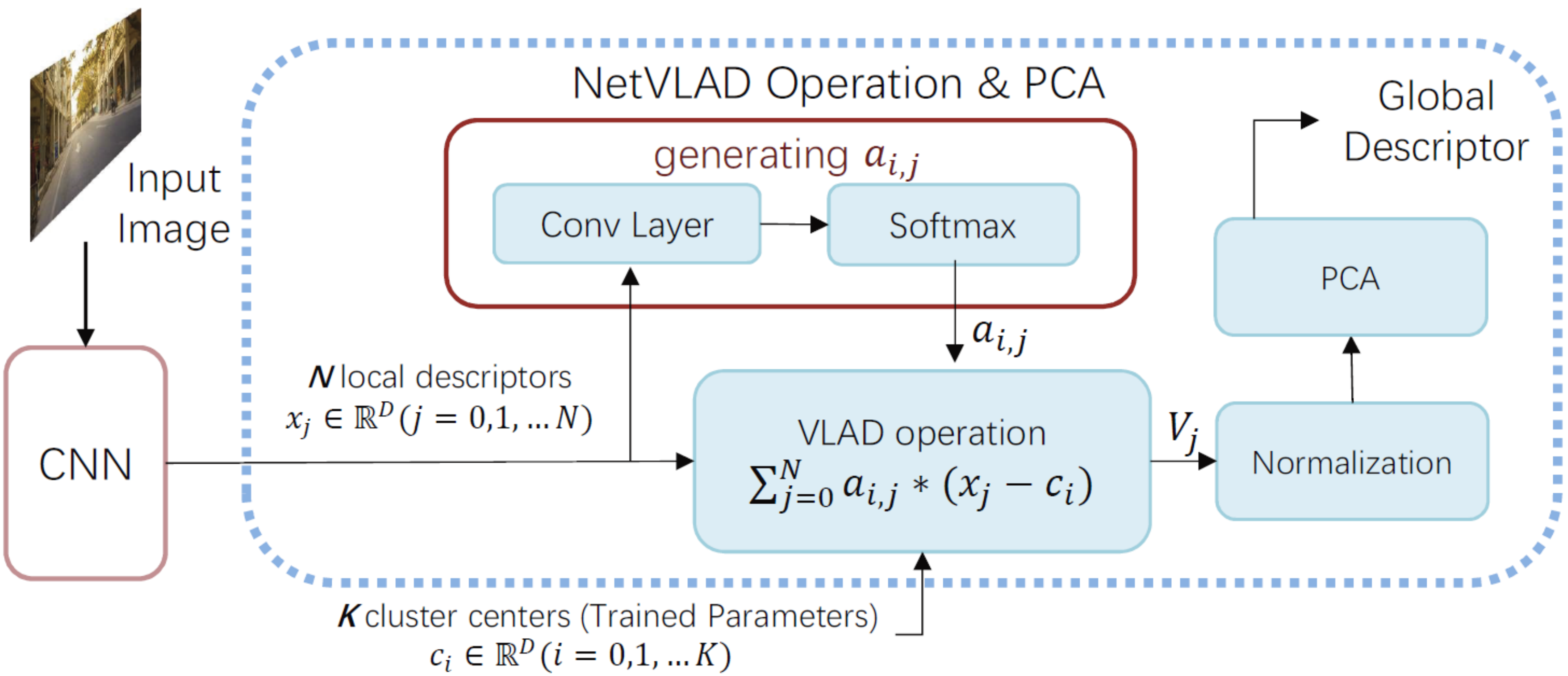




[1] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 815-823.
[2] Radenović F, Tolias G, Chum O. Fine-tuning CNN image retrieval with no human annotation[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(7): 1655-1668.

# CNN-based Place Recognition

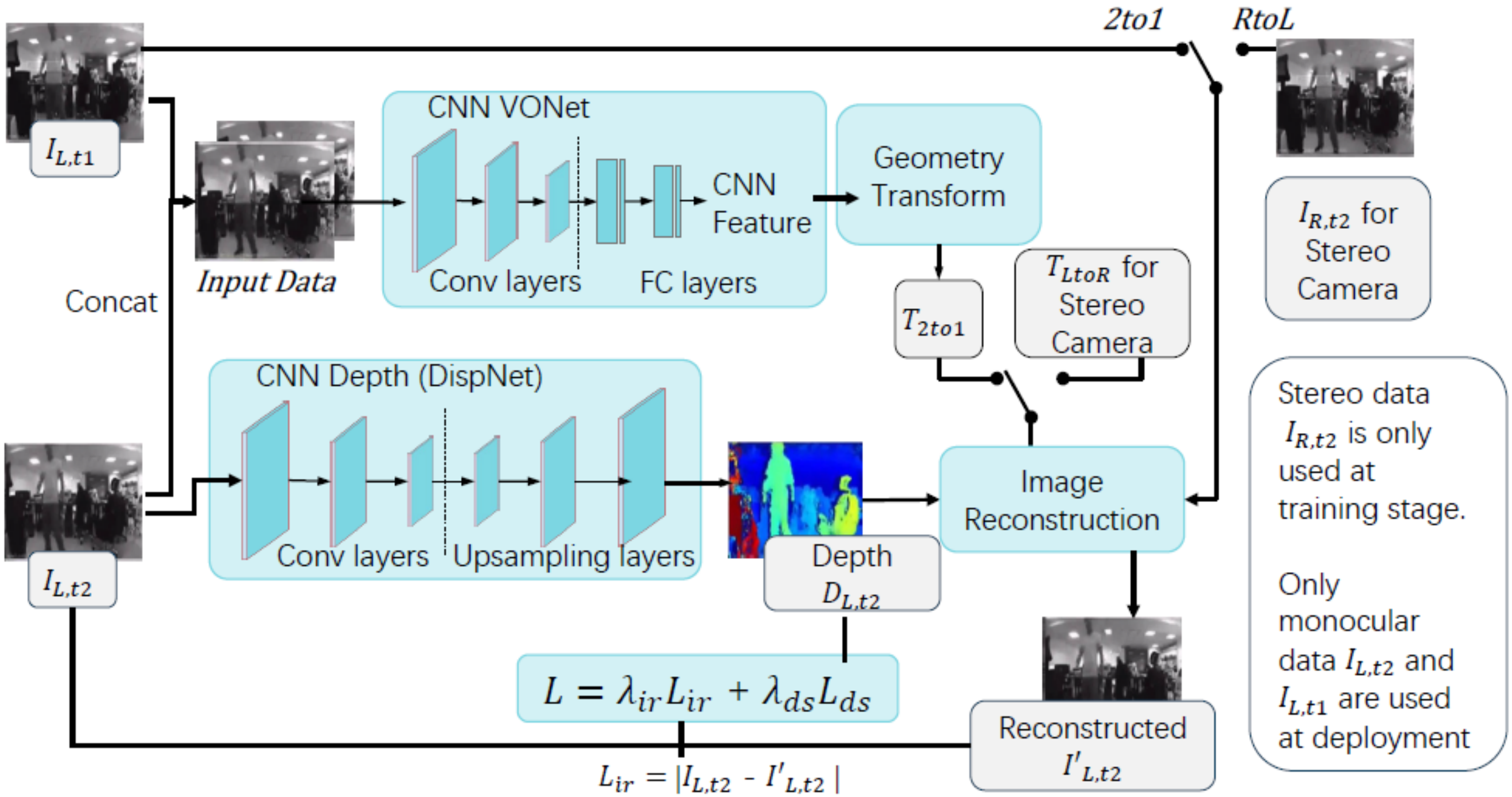
- **NetVLAD Pooling[1]**
  - Introduce the of idea Vector of Locally Aggregated Descriptors (VLAD) to CNN as a **pooling layer**.
  - Make the VLAD operation differentiable and thus trainable in the end-to-end CNN training process.




[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic,"NetVLAD: CNN architecture for weakly supervised place recognition,"in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2016, pp. 5297–5307.

# CNN-based Visual Odometry

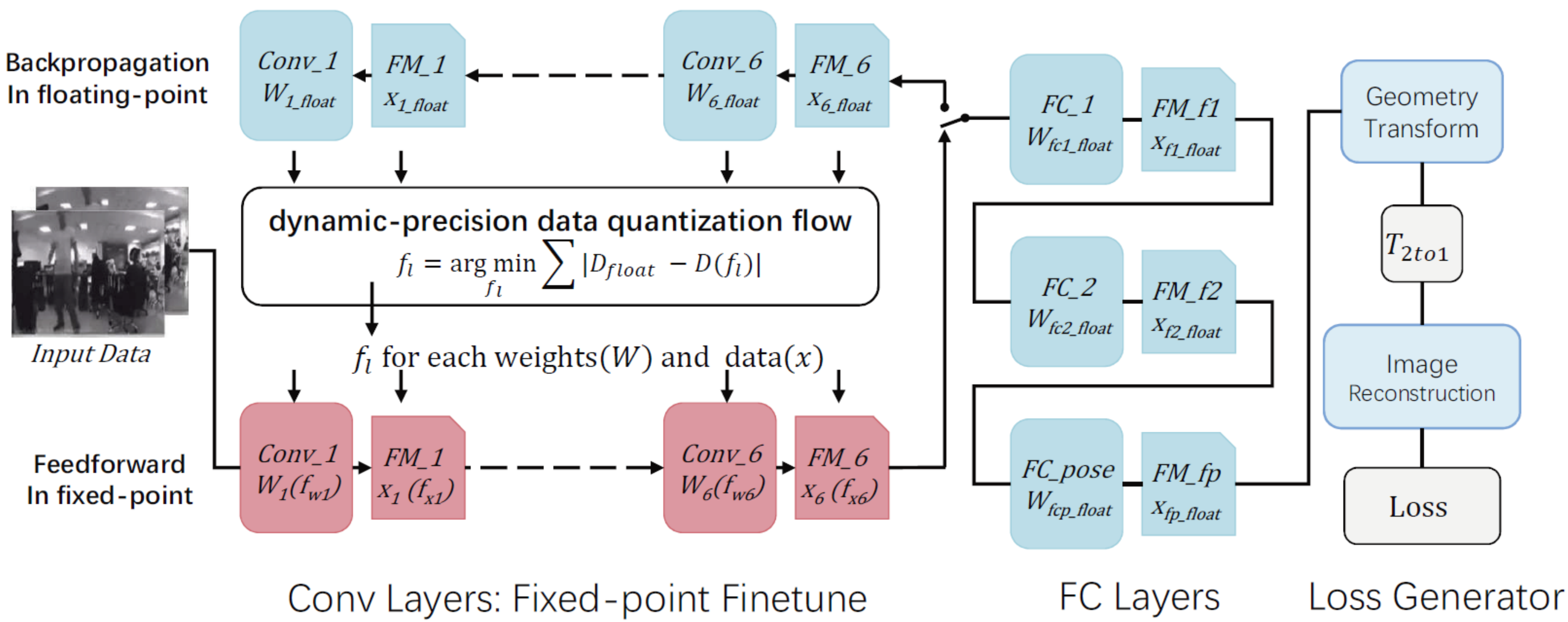
- **Depth-VO-Feat**[1]: A self-supervised end-to-end VO.
  - DispNet predicts the **depth** of the input frame.
  - VONet predicts the **VO results** of the two input frames.
  - VO results and depth can reconstruct $I_{L,t2}$ from $I_{L,t1}$ .
  - **Loss**: error between the reconstructed $I_{L,t2}$ and input $I_{L,t2}$ .




[1] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

# CNN-based Visual Odometry

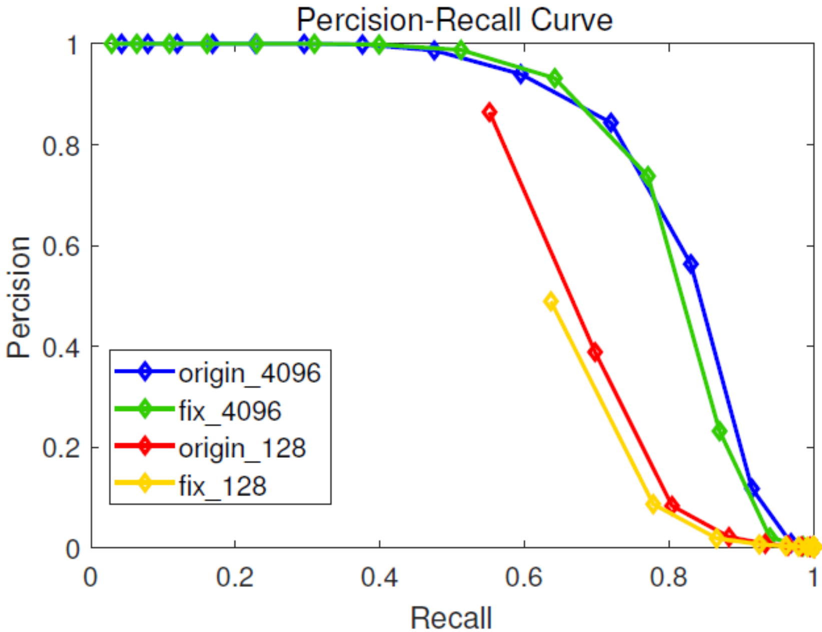
- Fixed-point fine-tune method
  - Floating-point number: loss generation and backpropagation.
  - Fixed-point number: feedforward of convolutional layers.
  - Data quantization flow converts the floating-point weights and intermediate featuremaps to fixed-point number.



# 8-bit Fixed-point Number for CNN

- The CNN layers are quantized with 8-bit fixed-point number on FPGA accelerator for DPR and VO

- **DPR:**
  - 8-bit direct quantization without fine-tune brings little accuracy loss.




- **VO:**
  - 8-bit for convolution layers (CONV), floating-point for fully connected layers (FC).

8-bit is better than floating-point

| | Seq. 09 | | Seq. 10 | |
|---|---|---|---|---|
| | translational drift error (%) | rotational drift error (°/100m) | translational drift error (%) | rotational drift error (°/100m) |
| Floating Point | 11.92 | 3.6 | 12.62 | 3.43 |
| 8-bit for CONV, floating-point for FC | 10.27 | 4.08 | 8.84 | 4.01 |
| 8-bit for CONV and FC layers | 13.27 | 5.27 | 14.75 | 7.78 |

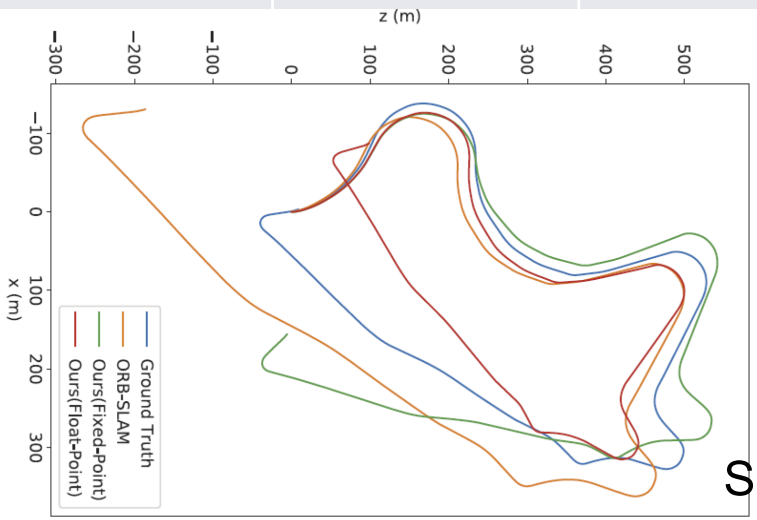
# VO Results on KITTI Dataset [1].

- CNN-based VO  runs on the embedded system (MPSoC ZCU102) in real-time.

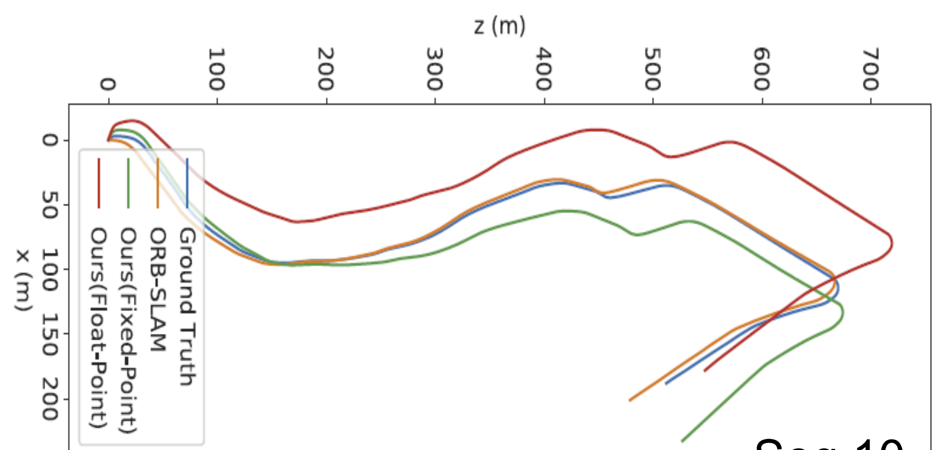
| | Quant. Strategy | Seq. 09 translational drift error (%) | Seq. 09 rotational drift error (°/100m) | Seq. 10 translational drift error (%) | Seq. 10 rotational drift error (°/100m) | Run rime (ms/frame) |
|---|---|---|---|---|---|---|
| Traditional method (ORB-SLAM[2]) | floating-point | 15.3 | 0.26 | 3.68 | 0.48 | 230 on embedded CPU |
| CNN based method (Depth-VO-Feat[3]) | floating-point | 11.92 | 3.6 | 12.62 | 3.43 | - |
| | 8-bit for CONV, floating-point for FC | 10.27 | 4.08 | 8.84 | 4.01 | 8 |

8-bit is better than floating-point

Real-time



Seq.09



Seq.10

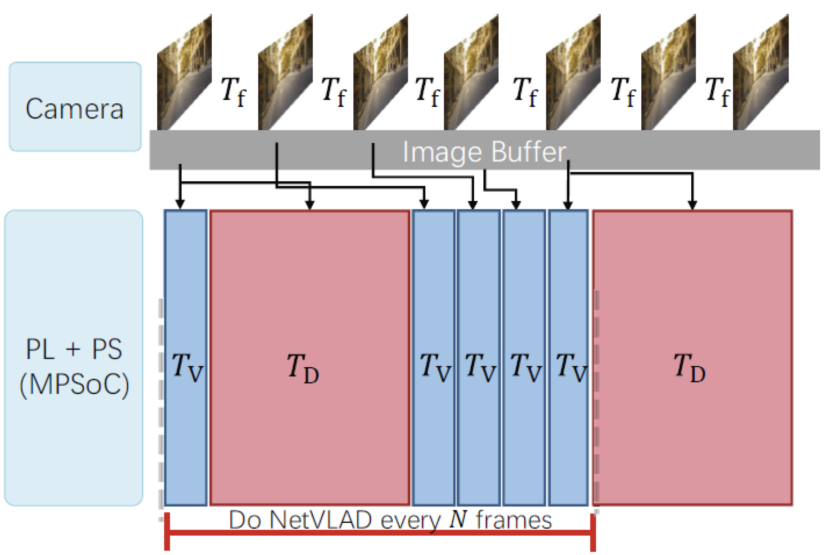

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," The International Journal of Robotics Research, vol. 32, no. 11, pp. 1231–1237, 2013.
[2] R. Mur-Artal and J. D. Tards, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," IEEE Transactions on Robotics, vol. 33, pp. 1255–1262, 2016.
[3] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

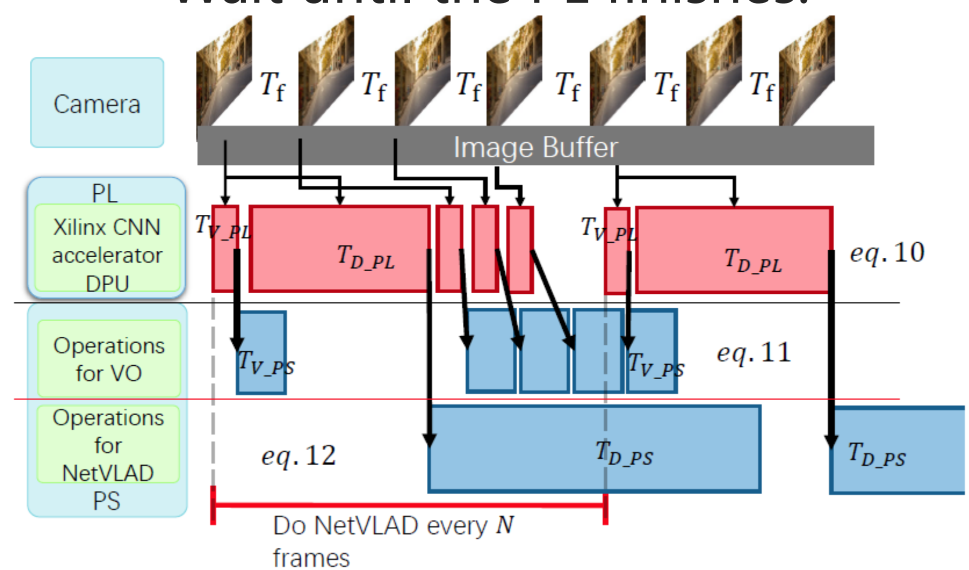
# Schedule 2 CNN Models on One Accelerator

- Two tasks: VO and DPR
- Two hardware parts: FPGA (Program Logic, PL) and CPU (Processing System, PS)

- **Serialize Scheduling**
  - Cache the input frame.
  - Wait until the whole process (PL&PS) finishes.

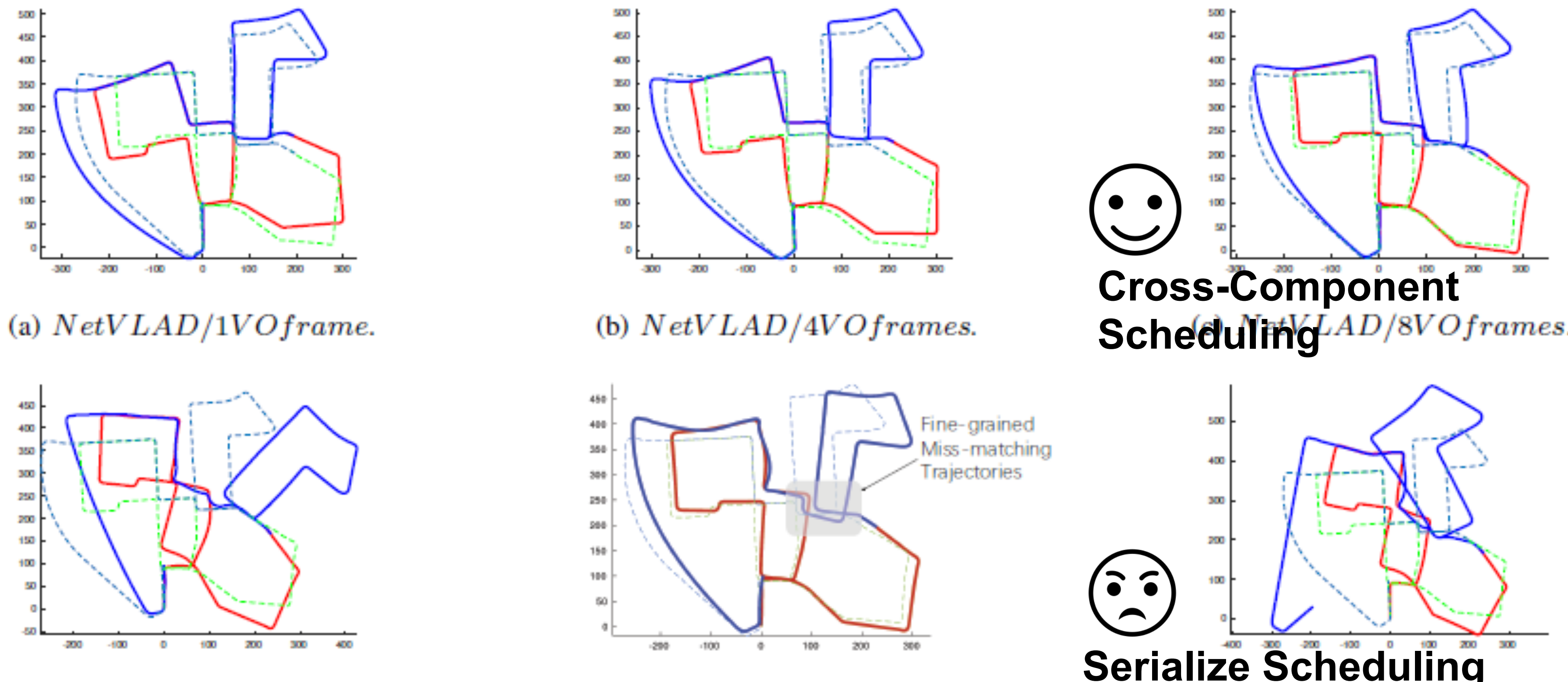
- **Cross-Component Scheduling**
  - The PS side supports multi-thread.
  - The PL side is single-thread.
  - Wait until the PL finishes.



# Cross-Component Scheduling Improves the DPR frequency

- The higher DPR frequency, the better DSLAM accuracy.
- Our cross-component pipeline improves DPR from every 14 input frames to 8 input frames.

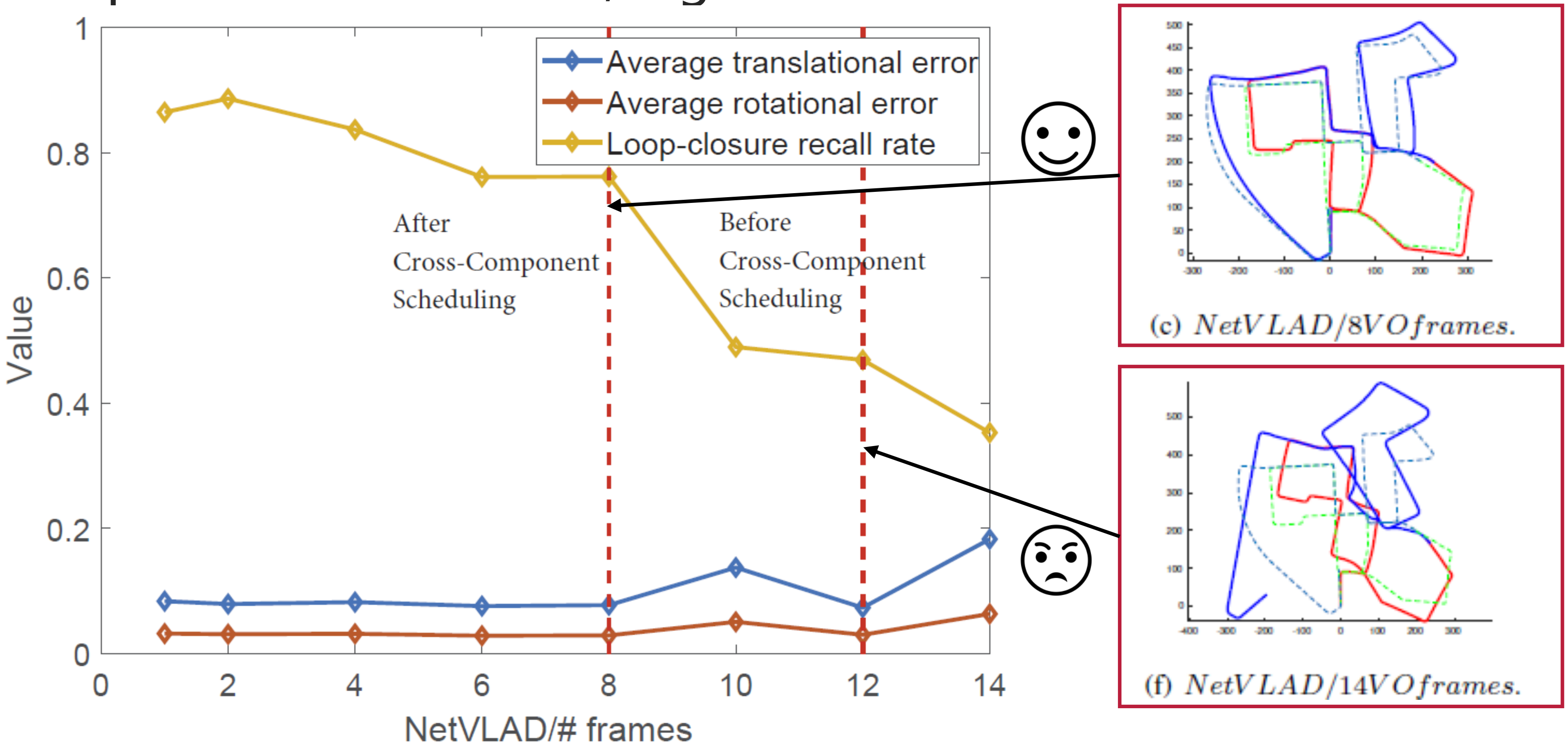



(a) $NetVLAD/1VOframe.$ (b) $NetVLAD/4VOframes.$ (c) $NetVLAD/8VOframes.$

d) $NetVLAD/10VOframes.$ (e) $NetVLAD/12VOframes.$ (f) $NetVLAD/14VOframes.$

# Cross-Component Scheduling Improves the DPR frequency

- ATE(Average translational error) , ARE(Average rotational error) indicate the error between estimated results and ground truth, lower is better.
- LCR (Loop-closure recall rate) indicates the success rate of loop-closure detection , higher is better.



# Conclusion

- Decentralized visual simultaneous localization and mapping (DSLAM) is the basic task of the multi-robot system.
- The two critical components in DSLAM, Decentralized Place Recognition (DPR) and Visual Odometry (VO) can benefit from CNN.
- Our **8-bit fixed-point quantization** is applicable in these novel CNN methods for robot applications.
- Our **cross-component scheduling** method can make 2 or more tasks use the same accelerator more efficiently, and thus improves the DPR frequency in DSLAM.
- Finally, we deploy the embedded real-time DSLAM system onto the MPSoC ZCU102 board.