独立行政法人　国立高等専門学校機構
大分工業高等専門学校
National Institute of Technology, Oita College

Amor Omnia Vincit

Facial Expression Recognition System Using DNN Accelerator with Multi-threading on FPGA

Takuto Ando[1]　Yusuke Inoue[2]

[1] *Electrical, Electronics Information Engineering Major,*

National Institute of Technology, Oita College Advanced Course

[2] *Department of Information Engineering, National Institute of Technology, Oita College*

2024 11/28 9th International Workshop on GPU Computing and AI (GCA'24)

Outline

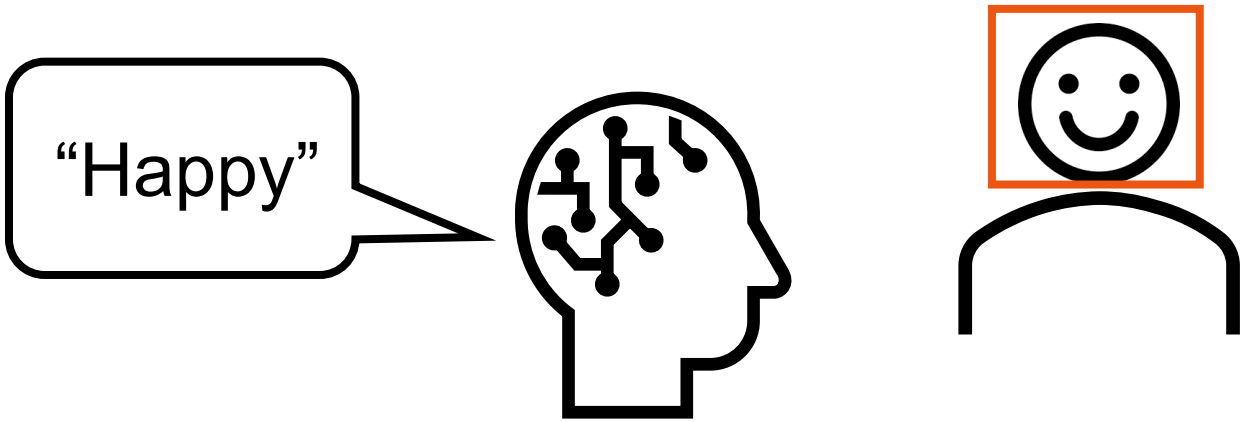
Outline

Facial expression and robot

Facial expression : Effective Non-Verbal Communication

- Human-to-Human : Conveys emotions and intentions
- Human-Computer : Enables intuitive interactions

▶ Applied to pet robots and medical robots



What is facial expression recognition?

Robots need to understand emotions

Ekman basic emotions classify them into six types [1]

Ekman's six basic emotions

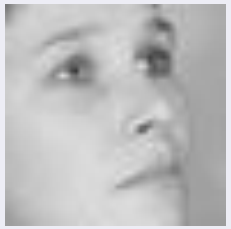
Anger

Disgust

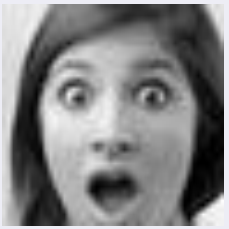
Fear

Happy

Sad



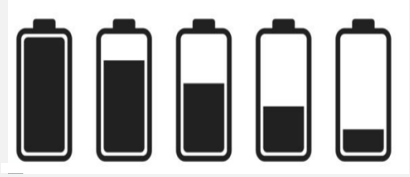
Surprise



[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," J. Pers. Soc. Psychol., vol. 17, no. 2, pp. 124–129, 1971.

Implementation on robot

Installation on battery-powered robots



- Low-power processing for longer operation
- DNN recognition needs a high-performance unit (e.g., GPU)

☹ GPU provides high performance <u>BUT high-power consumption</u>

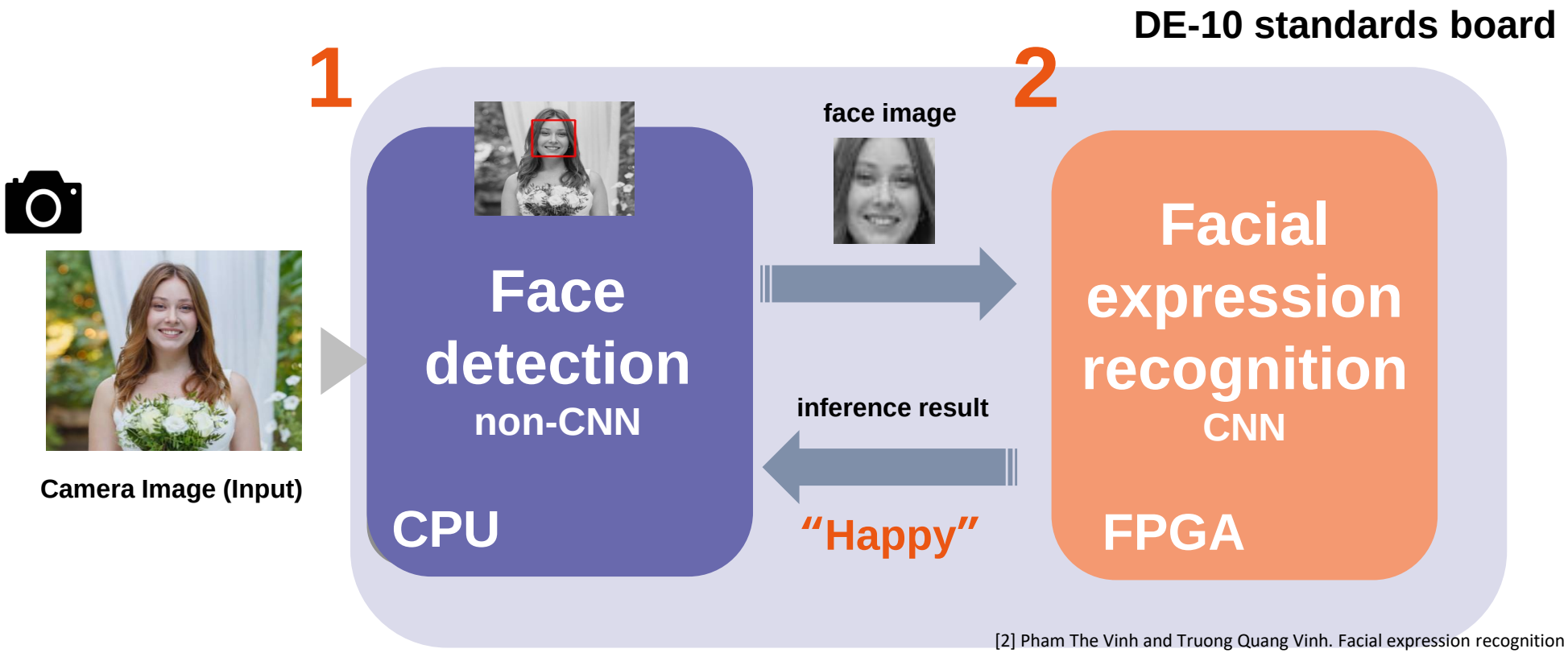
☺ **FPGA Implementation** : Balanced Solution

low power consumption and high computing performance

FPGA board
Xilinx Kria KV260

Previous work

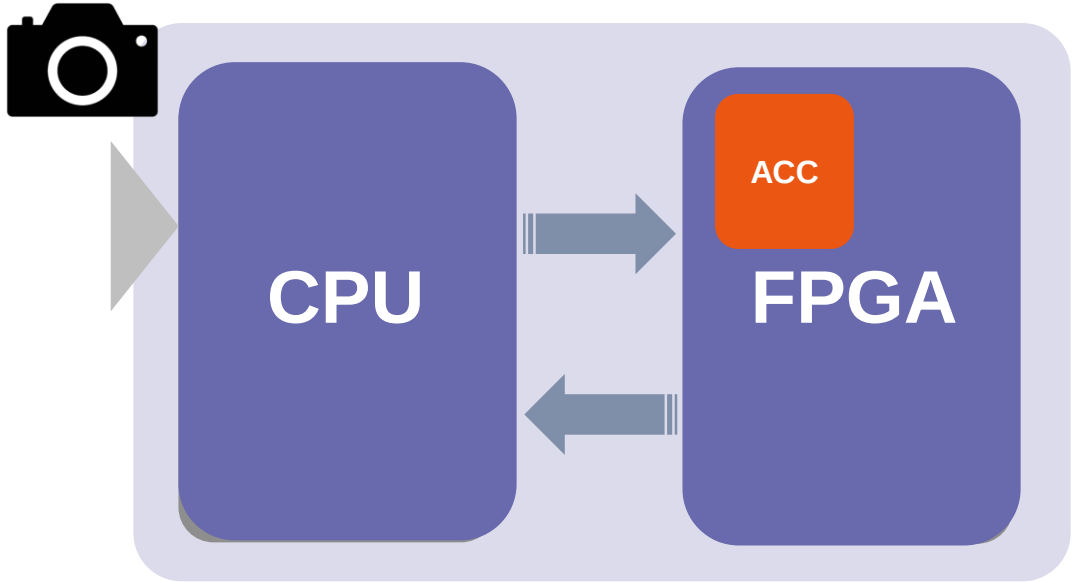
Vinh et al. implemented a facial expression recognition system using an SoC FPGA [2]



[2] Pham The Vinh and Truong Quang Vinh. Facial expression recognition system on soc fpga. In 2019 International Symposium on Electrical and Electronics Engineering (ISEE), pp. 1–4. IEEE, 2019.

In this work

Previous work

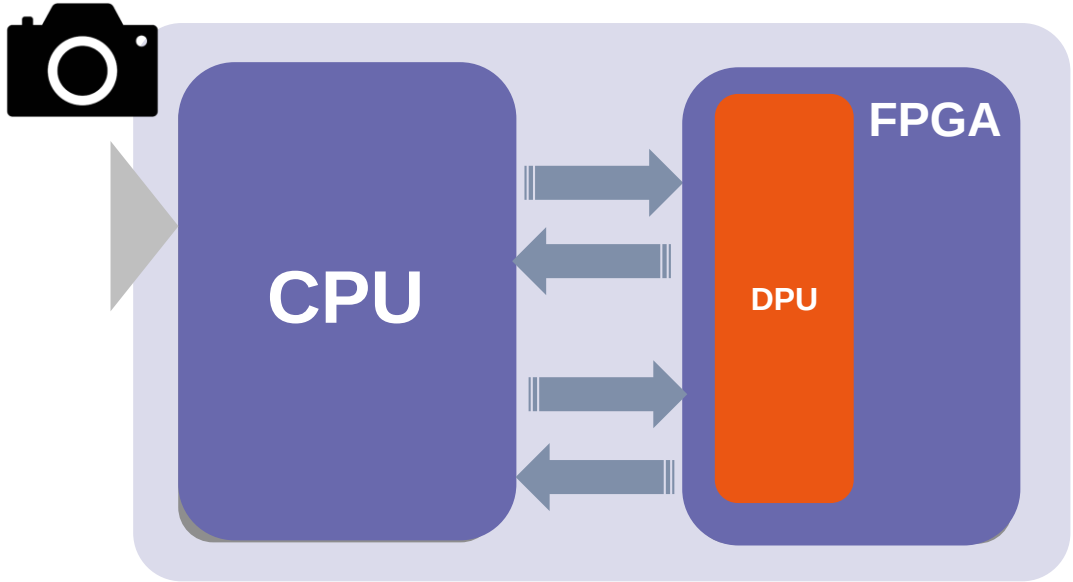


FD (non-DNN)

FER (DNN)

Overall control

Our work



Overall control

FD (DNN)

FER (DNN)

FD : Face Detection
FER : Facial Expression Recognition

ACC : accelerator
DPU : Deep learning Processing Unit

Objectives

Running Two DNN Models on the Same DPU

- Improve DPU utilization efficiency with multi-threading
- Achieve high throughput and low power consumption

System Implementation and Evaluation

- Offloaded two DNN inferences to FPGA
- Face detection and facial expression recognition

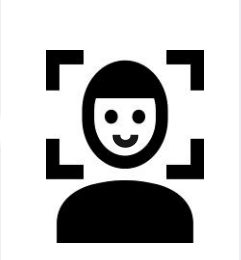
Outline

Two DNN models

01 Face detection : DenseBox

Input : 640 x 460 x 3

Output : Coordinates of the face region (x, y, w, h)



02 Facial expression recognition : CNN

Input : 48 x 48 x 1

Output : Label of the expression class (7 categories)

Face detection model

Dense Box [3] : Face detection model provided by Xilinx

- Lightweight and simple network
- The Wider Face dataset [4] was used for training

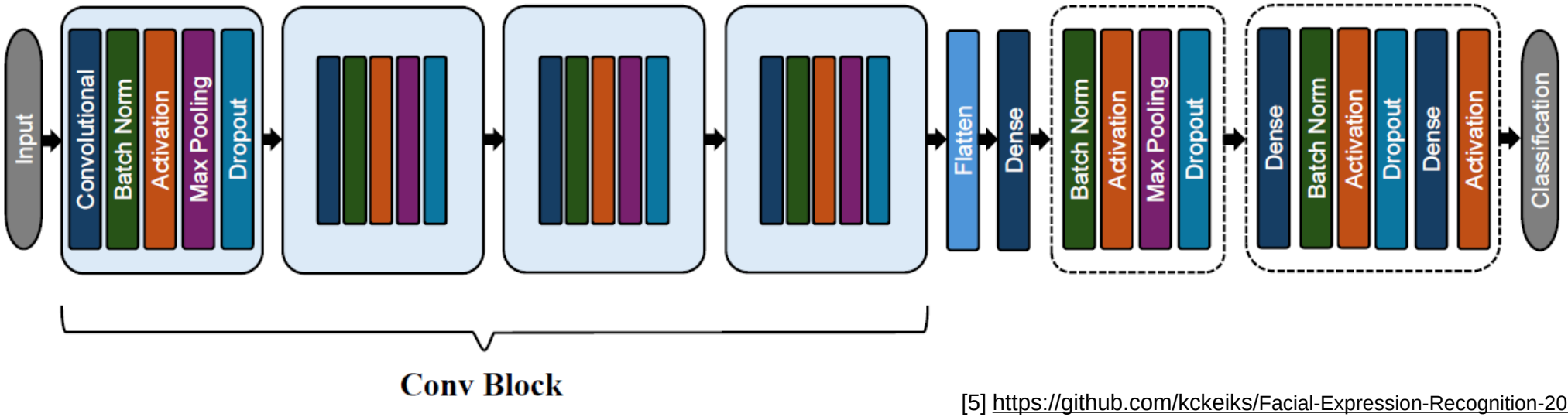
Face detection using Dense box example [3]

[3] Vitis AI Library User Guide UG1354 (v3.5) June 29, 2023
[4] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5525–5533, 2016.

Facial expression recognition model

Guarniz's CNN architecture [5]

- Feature extraction with 4 repeated blocks (convolution, batch normalization, activation, pooling, dropout layers)
- Trained using FER-2013 dataset



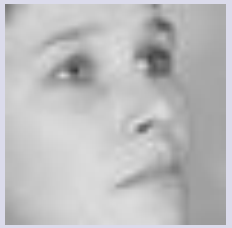
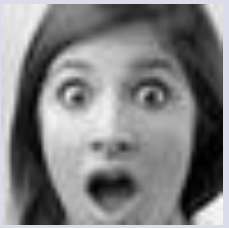
[5] https://github.com/kckeiks/Facial-Expression-Recognition-2018/tree/master (Accessed on 2024,9/1)

FER-2013 [6]

7 facial expression labels (Ekman's basic emotions + "Neutral"):

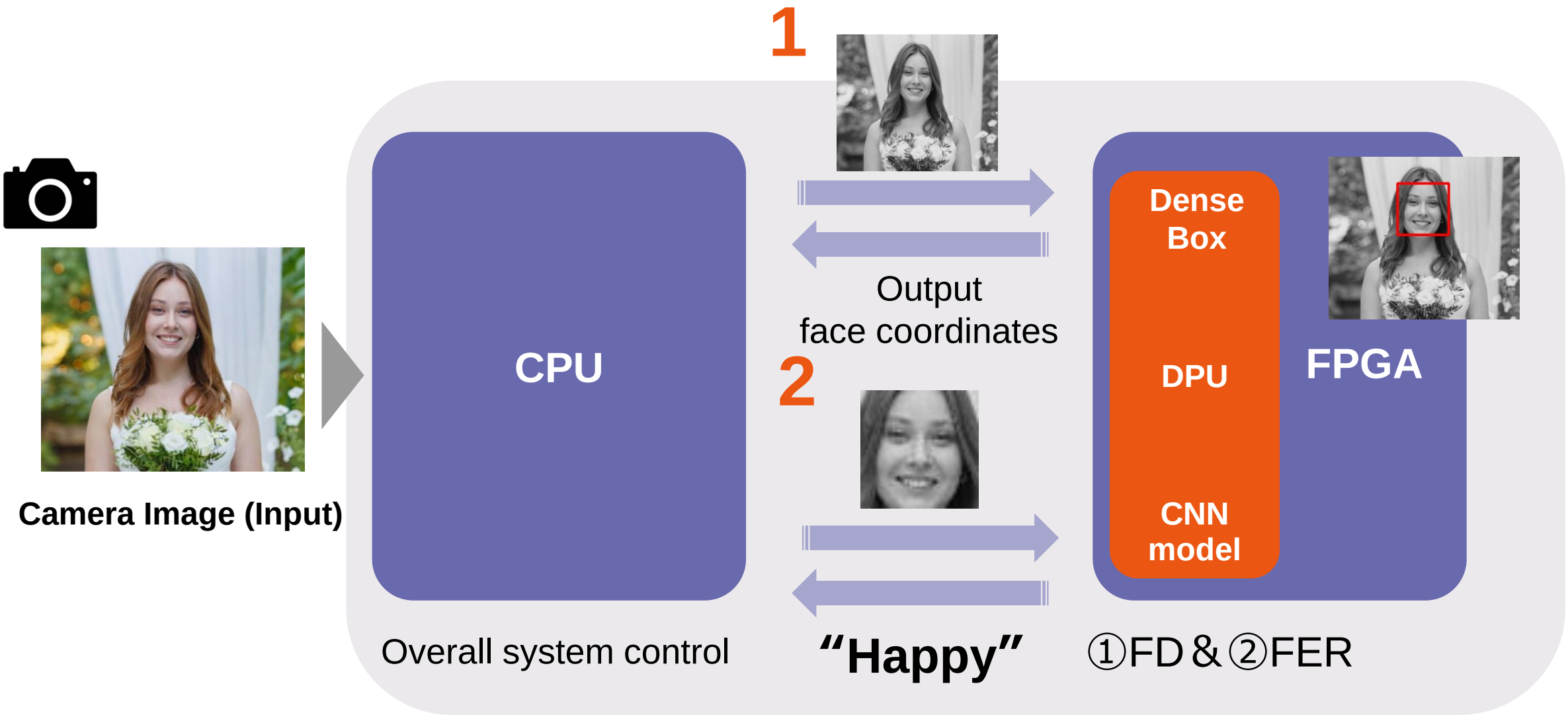
- **Training set**: Approx. 27,000 images **Test set**: Approx. 3,500 images

Face images: 48x48 pixels

Anger	Disgust	Fear	Happy	Sadness	Surprise	Neutral
						

[6] Carrier, P.L.; Courville, A.; Goodfellow, I.J.; Mirza, M.; Bengio, Y. FER-2013 Face Database; Universit de Montral: Montreal, QC, Canada, (2013).

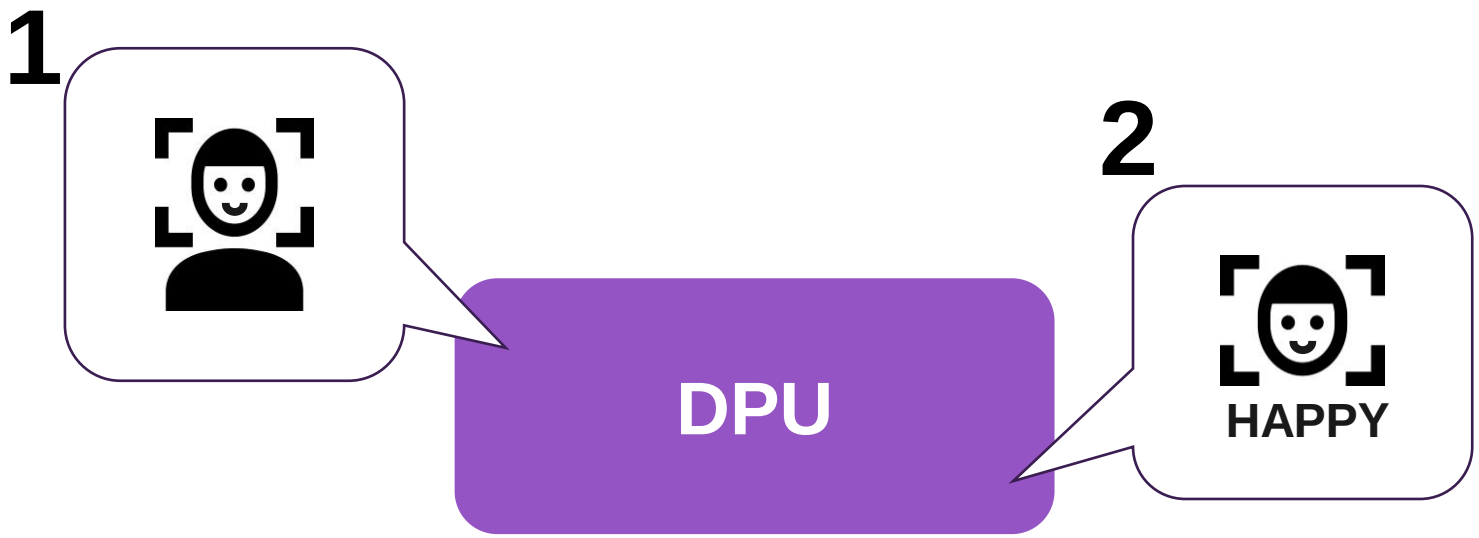
Hardware configuration



What is “DPU” ?

Deep learning Processing Unit

- Implemented on FPGA as a CNN accelerator
- Provided by Xilinx
- Multiple DNN models executed in time division on the same DPU



What is “DPU” ?

Purposeful use is possible

- Select an architecture from B512 to B4096 for each application
- The higher the number, the higher the performance

Processing performance

FPGA resources for different DPU sizes

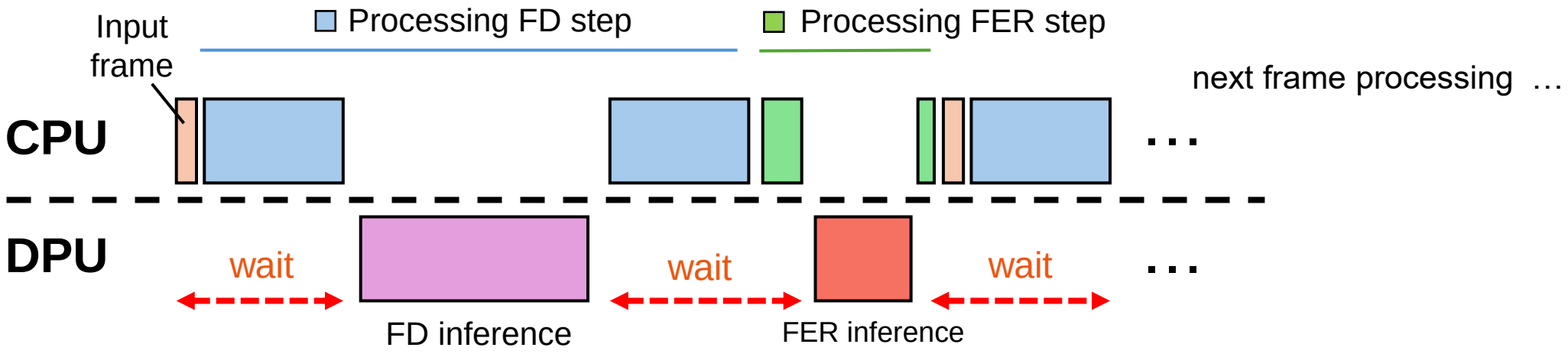
DPU Architecture	LUT	Register	Block RAM	DSP
B512	**26922**	**34543**	**72**	**118**
B1024	34074	48057	104	230
B2304	42127	68829	165	438
B4096	52161	98249	255	710

Lightweight circuits

Multi-threading strategy

Utilization efficiency of the DPU in single-threading

- The DPU has idle time until it is given instructions
- This method does not utilize the DPU efficiently



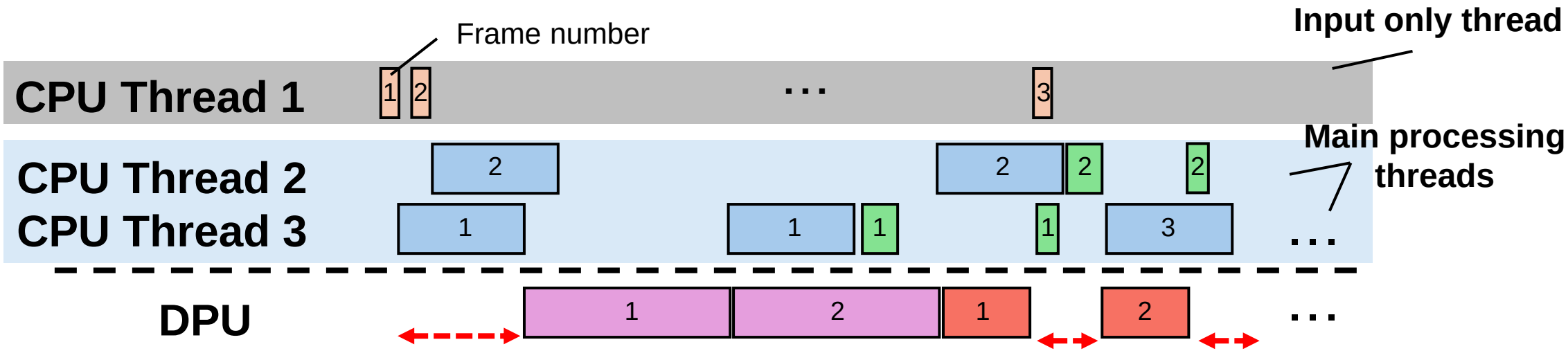
The idle time is long

FD : Face Detection
FER : Facial Expression Recognition

Multi-threading strategy

Utilization efficiency of the DPU in multi-threading

- Increased frequency of tasks assigned to DPU
- Reduced waiting time, enabling more efficient DPU operation



The idle time is short

Outline

Experimental objectives

Assess the effectiveness of offloading DNN inference to the DPU

Comparison with previous work

- Recognition performance evaluation for each DNN model
  - Recognition accuracy
  - Processing time
- System evaluation : overall system performance

Evaluation board

Integrates a CPU and FPGA on the same chip

Target board : Xilinx Kria KV260

CPU : ARM Cortex-A53

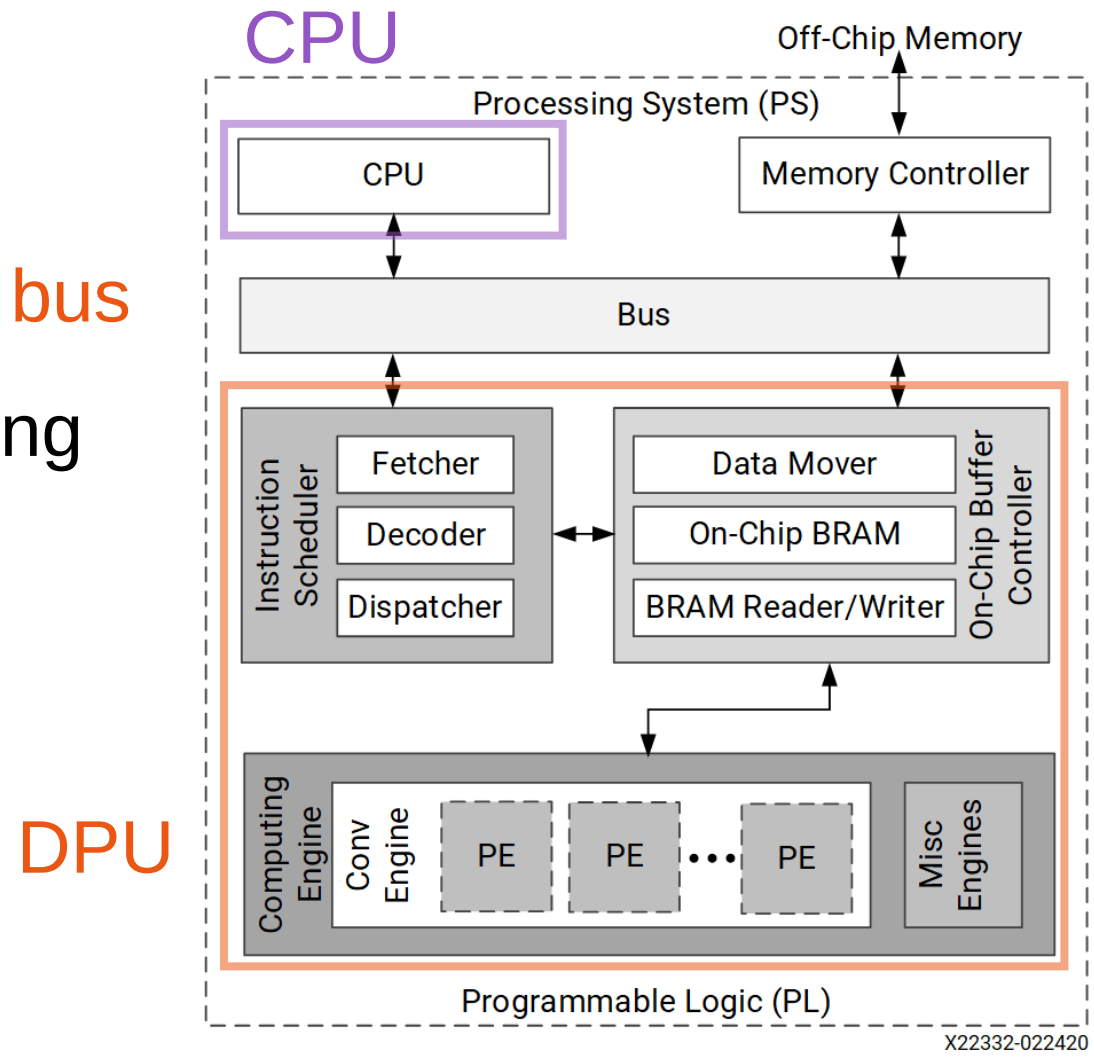
FPGA : Xilinx UltraScale+

DDR memory : 4 GB

Operating frequency : 1.3 GHz

Hardware architecture

- CPU and DPU communicate via AXI bus
- DPU operation is controlled by fetching instructions from off-chip memory



Hardware Architecture Overview [7]

[7] https://docs.amd.com/r/en-US/pg338-dpu/Hardware Architecture

Evaluate each inference

01 Face detection

Using AFW dataset (401 images) [8]

Compare AP and processing time per image

02 Facial expression recognition

Using the FER dataset (3,589 images)

Compare accuracy and processing time per image

[8] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886, 2012.

Evaluate each inference

01 Face detection

Using AFW dataset (401 images) [8]

Compare AP and processing time per image

02 Facial expression recognition

Using the FER dataset (3,589 images)

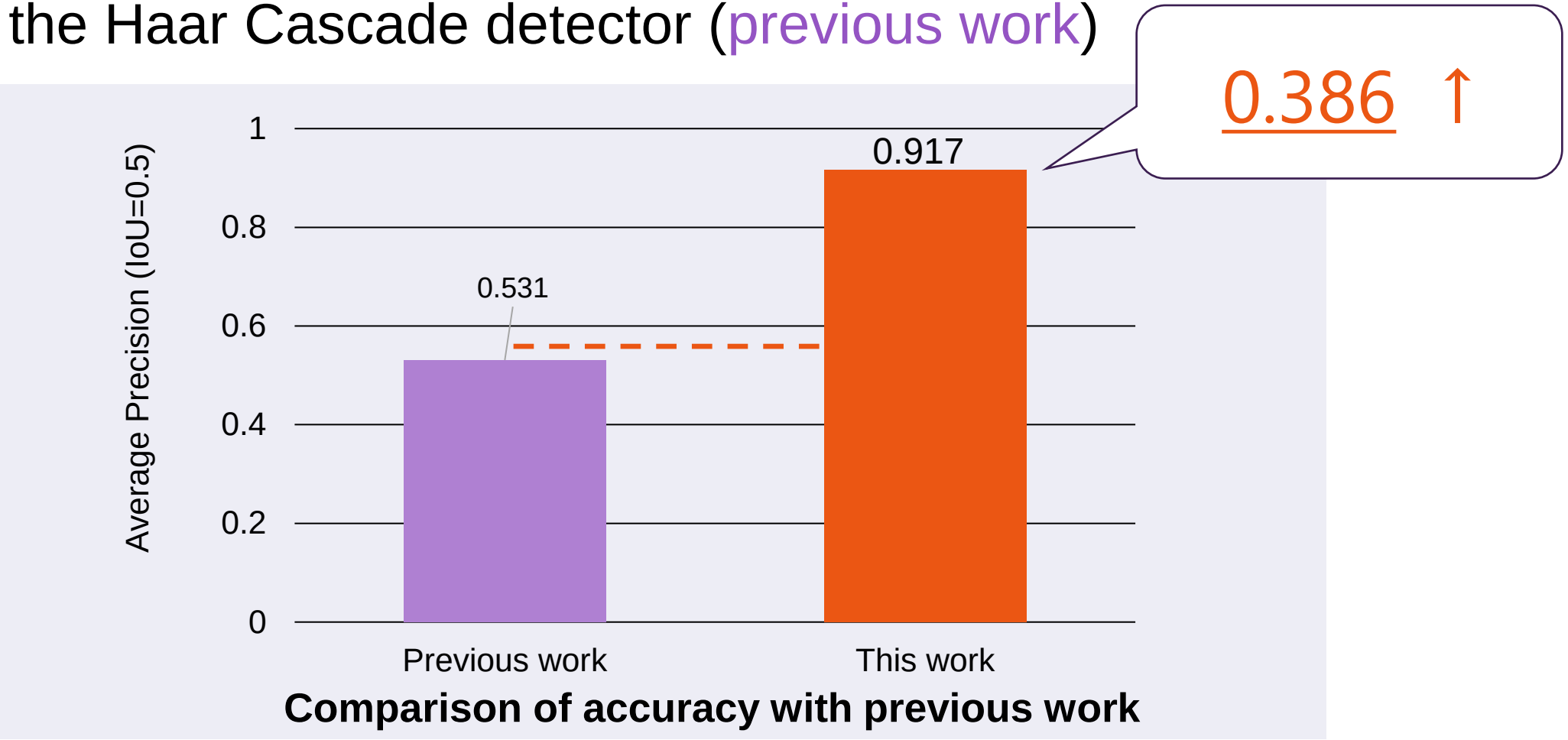
Compare accuracy and processing time per image

[8] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886, 2012.

Accuracy of face detection

Comparison of the DNN model (this work)

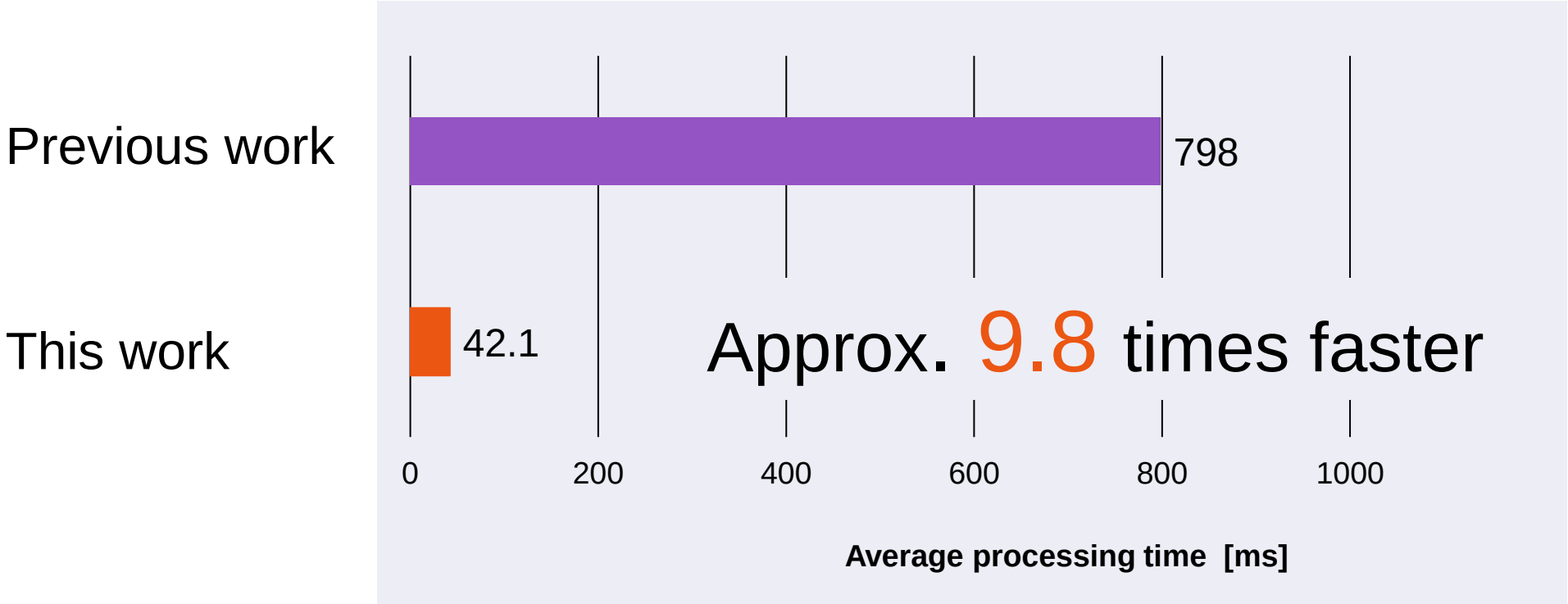
with the Haar Cascade detector (previous work)



Processing time face detection

Comparison of the DNN model (this work)

with the Haar Cascade detector (previous work)



Comparison of average processing time with previous work

Evaluate each inference

01 Face detection

Using AFW dataset (401 images)

Compare AP and processing time per image

02 Facial expression recognition

Using the FER dataset (3,589 images)

Compare accuracy and processing time per image

Facial expression recognition results

[3] L. Pham, T. H. Vu and T. A. Tran, "Facial Expression Recognition Using Residual Masking Network," *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 4513-4519.

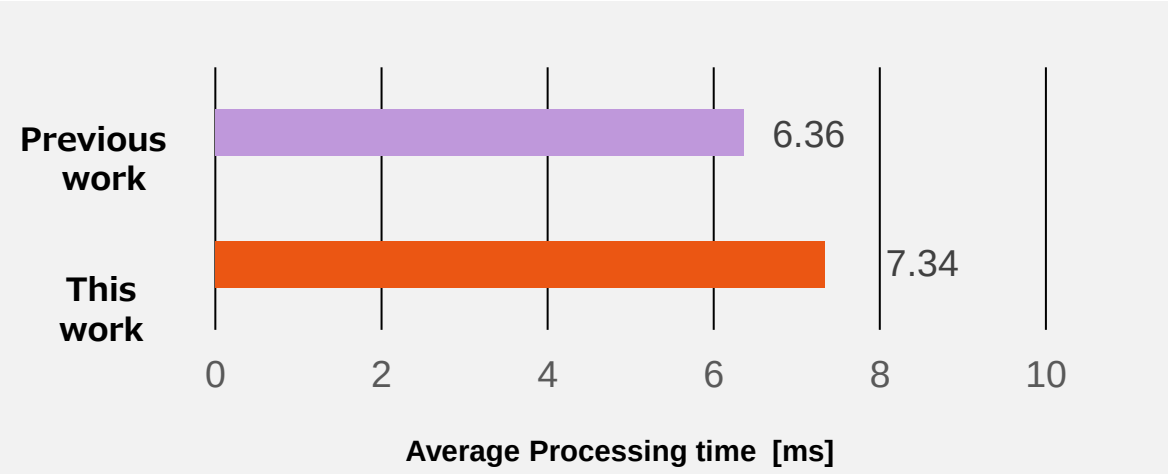
Comparison with CNN model from previous work

1. Recognition accuracy [%]

Comparison of accuracy with previous work

- Slightly higher than the previous work
- Lower than SOTA but acceptable

2. Processing time / image [ms]



Comparison of processing time with previous work

- Previous work is about 1 ms superior
- Almost ignorable due to face detection time

Overall system comparison

Evaluate throughput and power consumption

- **Throughput**
  - Measure overall system throughput
- **Power consumption** (SoC)
  - Compare idle state (7.8W) with system runtime
  - Measure using KETOTEK KTEM02 connected to the board's power socket

KETOTEK KTEM02 digital energy meter

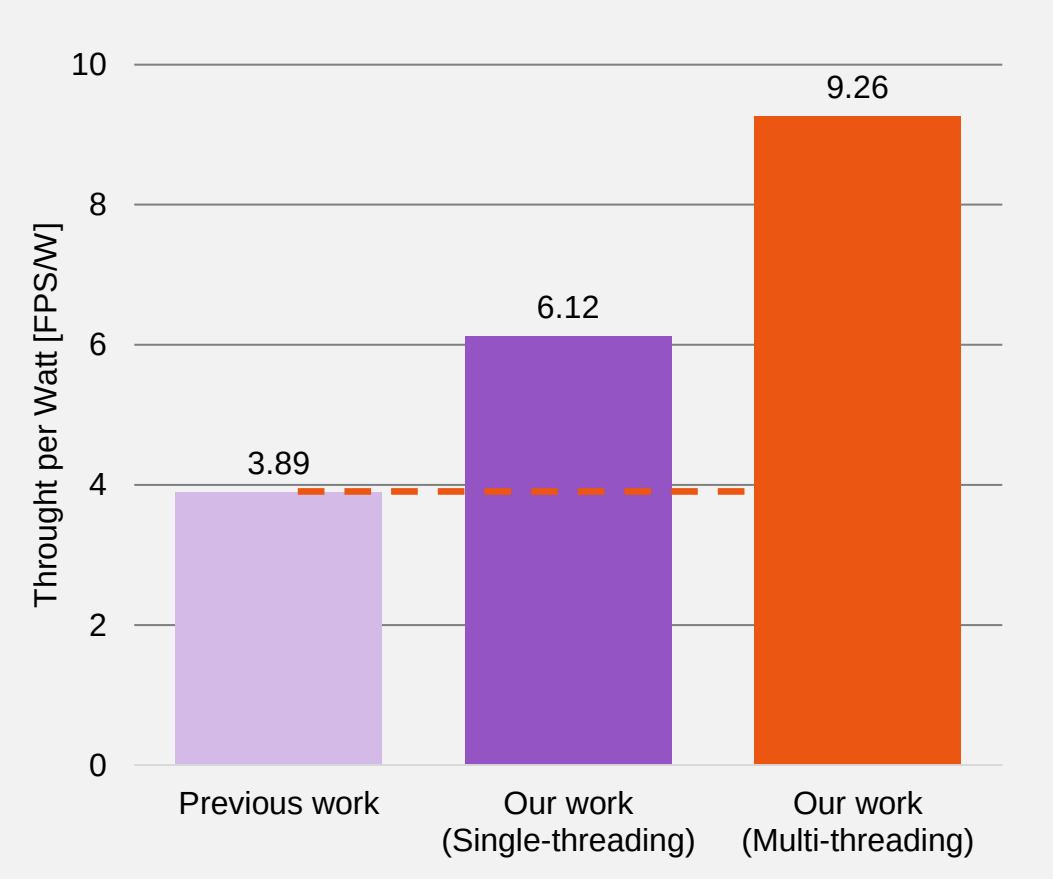
Overall system comparison

Evaluate throughput and power consumption

Throughput

Previous work	:	11.67 FPS
Our work (single-threading)	:	14.69 FPS
Our work (multi-threading)	:	25.00 FPS

Realize real-time performance



Comparison of throughput per power consumption

Compare circuit size

Compare FPGA resource utilization with previous work

This system is slightly larger

Comparison of FPGA resource utilization

	ALM or LUT	DSP	BRAM
Previous work (Intel: ALM)	22,465	112	44
Our work (Xilinx: LUT)	27,023	118	12

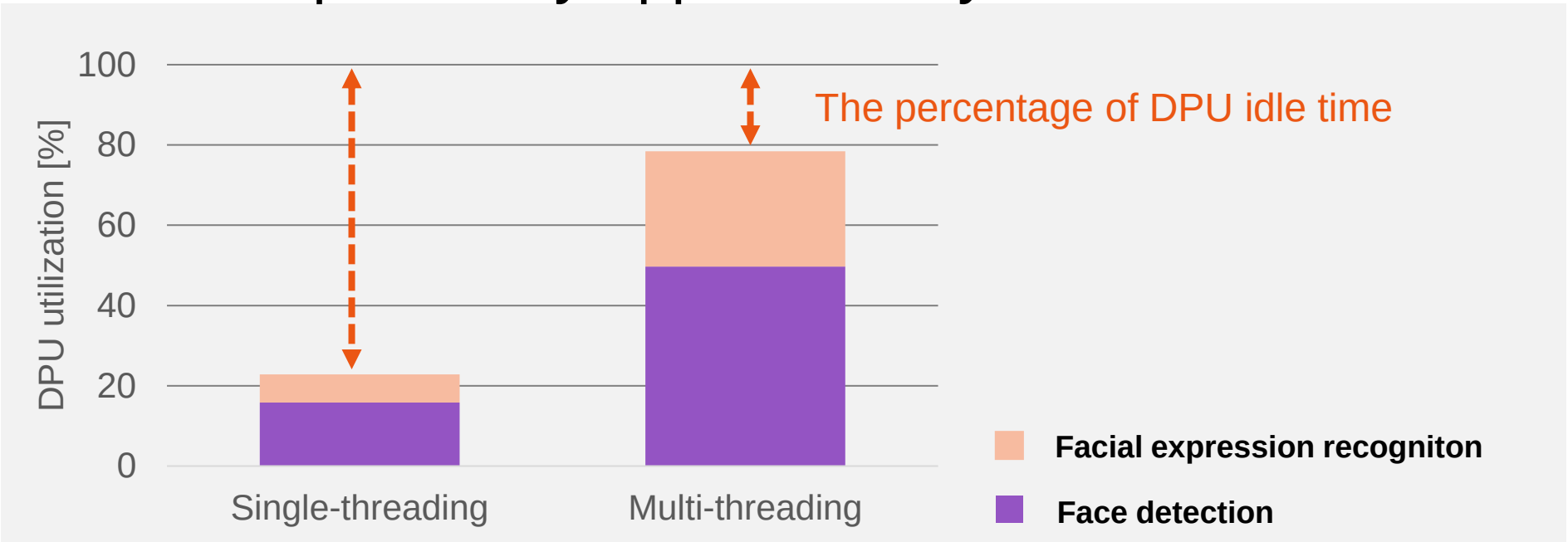
Two DNN inferences could be performed
without a significant increase in circuit size

Verify utilization efficiency of DPU

Compare DPU utilization by threading

Calculate DPU processing time as a percentage of system runtime

DPU utilization improved by approximately 3.43x



DPU utilization comparison by threading

Outline

01	Introduction	Problem Background and objectives
02	Proposed method	Explanation of the FER system on DPU
03	Experimental evaluation	Evaluate by comparing performance with the previous work
04	Disscution	Investigate optimal DPU size and frequency
05	Conclusion	Conclusion of this presentation and future work

Analysis of optimum operating frequency

Investigate optimal DPU operating frequency

The throughput improvement rate is low even

when the operating frequency exceeds 400 MHz

Throughput per power consumption by frequency

Frequency	Throughput [FPS]	Peak Power [W]	Power [W]	Throughput/Power [FPS / W]
600	27.78	11.7	2.3	12.08
500	25.00	11.6	2.2	11.36
400	**25.00**	**11.4**	**2.0**	**12.50**
300	21.74	11.2	1.8	12.08

Investigated optimal DPU size

FPGA resources and performance

Comparison made between B512 and larger DPU at 400 MHz

For larger DPU, throughput is not worth the circuit size

Comparison of FPGA resources and performance with different DPU sizes

Size	FPGA resource			Thread	Throughput [FPS]	Power [W]
	LUTs	DSPs	BRAMs			
512	**27,023**	**118**	**12.0**	1	14.69	2.4
				2	**25.00**	**2.7**
1024	34,593	230	44.0	1	19.21	2.5
				2	27.74	3.1
2034	41,861	438	60.5	1	20.80	2.9
				2	27.75	3.2
4096	51,561	710	82.5	1	23.77	3.2
				2	27.74	3.5

9.26 FPS/W

Outline

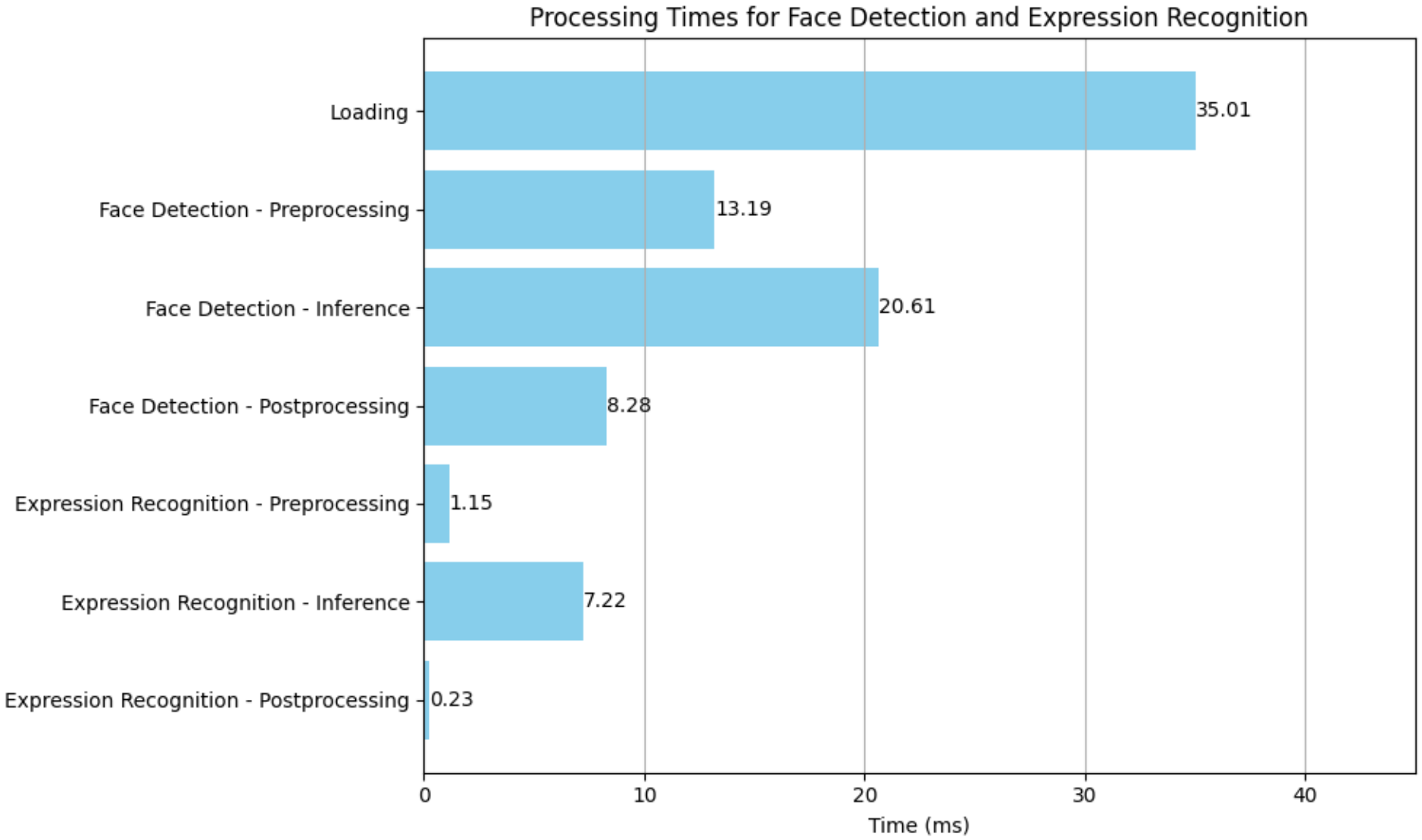
01	Introduction	Problem Background and objectives
02	Proposed method	Explanation of the FER system on DPU
03	Experimental evaluation	Evaluate by comparing performance with the previous work
04	Disscution	Investigate optimal DPU size and frequency
05	Conclusion	Conclusion of this presentation and future work

Conclusion

- We implemented facial expression recognition system on DPU
- We utilized a systolic array accelerator for time-division inference of two DNNs on the same DPU
- We proposed a multi-threaded system to improve throughput and DPU utilization efficiency
- **Future work :** Reduce power consumption, optimize processing for real-world applications (e.g., face detection every few frames)

appendix

Processing Times for System



Privious work challengings

Low accuracy with Haar Cascade detector running on CPU

Very lightweight ○

Not robust to detect oblique or sideways faces ✕

Very sensitive to lighting ✕

Poor facial expression recognition accuracy

due to no proper face detection

Quantization & Compliation

Quantization : Converts 32-bit floating type → 8-bit integer type

Reduces the number of hardware operations

Compilation: Converted to a format executable by DPU

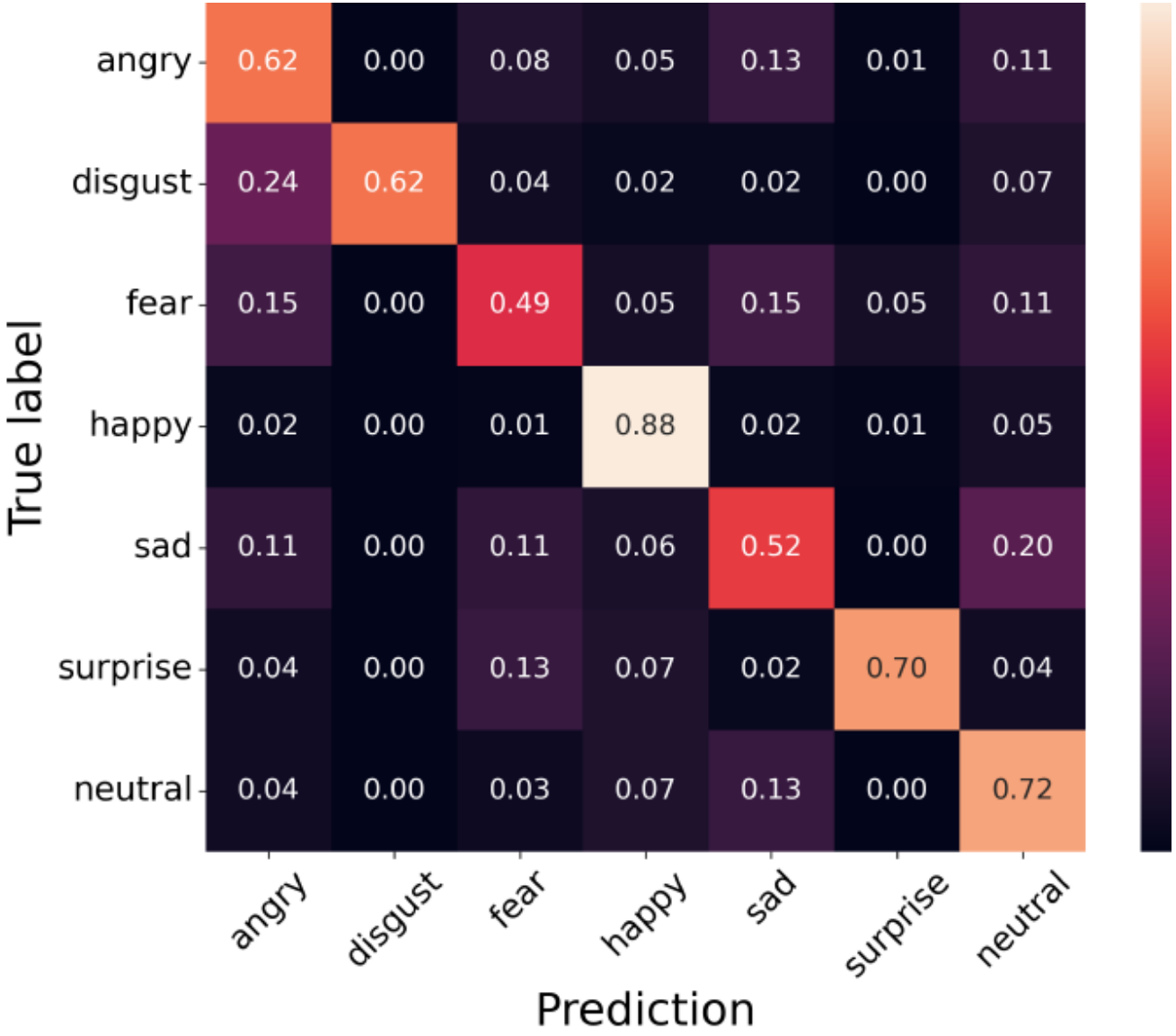
Accuracy by Expression Classes

Confusion matrix

Fear class △

Sad class △

Otherwise, high accuracy



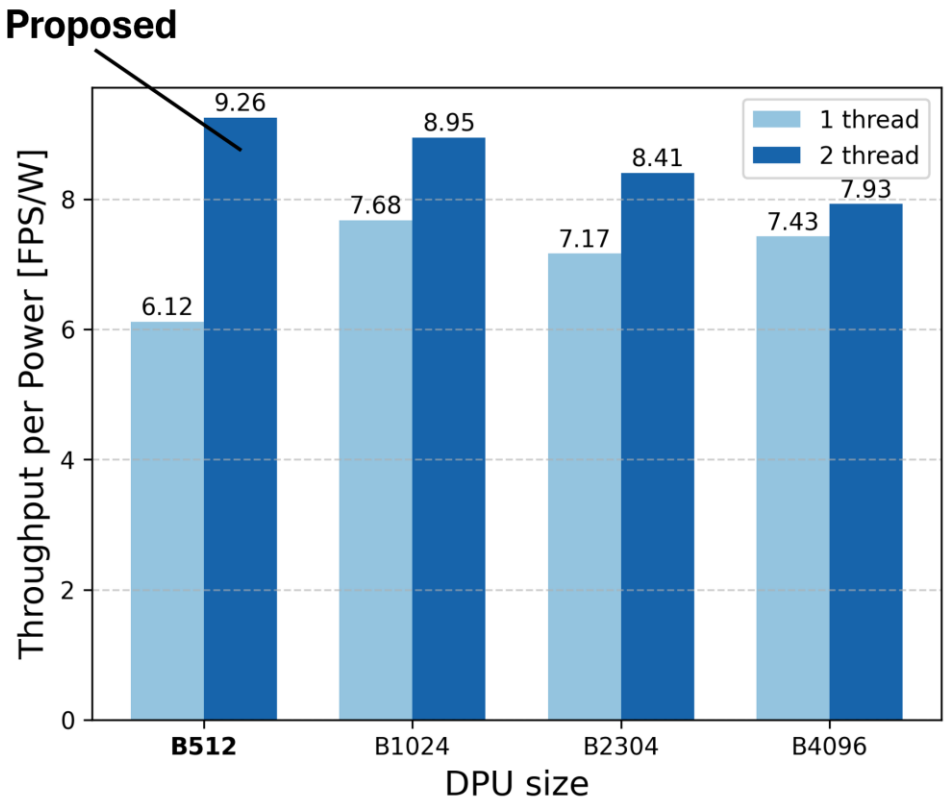
Compare with other sizes of the DPU

Investigated throughput per power consumption

Comparison made between B512 and larger DPU at 400 MHz

B512 multi-threading execution

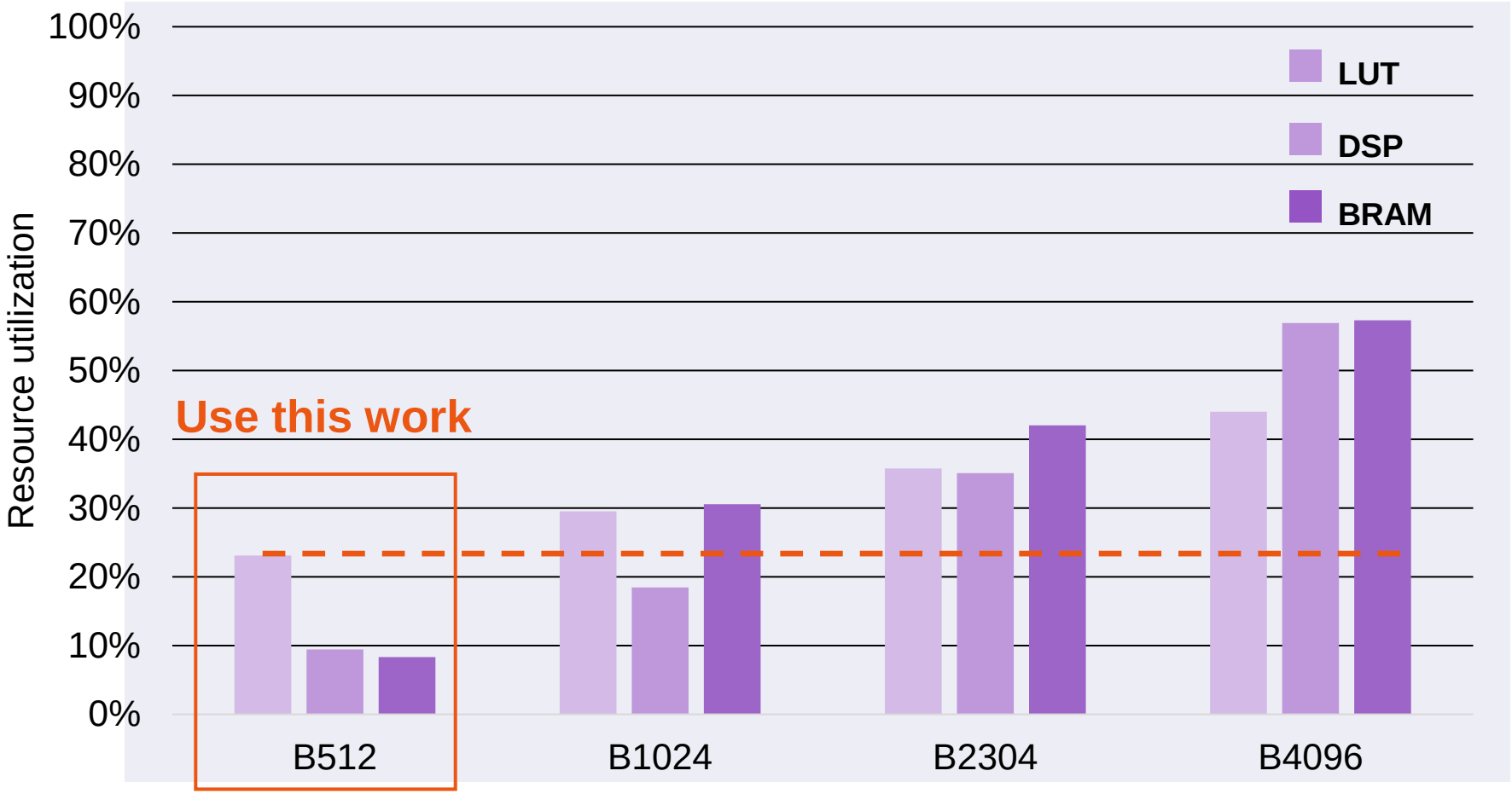
achieves highest efficiency of 9.26 FPS/W



Throughput per power consumption for hardware with each DPU

Evaluation by DPU size

FPGA resource utilization ：Kria KV260



Comparison of circuit size embedding each DPU

スライド素材

Outline

01	Introduction	Problem Background and objectives
02	Proposed method	Explanation of the FER system on DPU
03	Experimental evaluation	Evaluate by comparing performance with the previous work
04	Disscution	Investigate optimal DPU size and frequency
05	Conclusion	Conclusion of this presentation and future work