IDEC Chip Design Contest

DT-CNN: Dilated and Transposed Convolutional Neural Network Accelerator for Real-time Image Segmentation on Mobile Devices

Dongseok Im, Donghyeon Han, Sungpill Choi, Sanghoon Kang, and Hoi-Jun Yoo

School of Electrical Engineering
Korea Advanced Institute of Science and Technology
Daejeon, South Korea

Motivation & Introduction

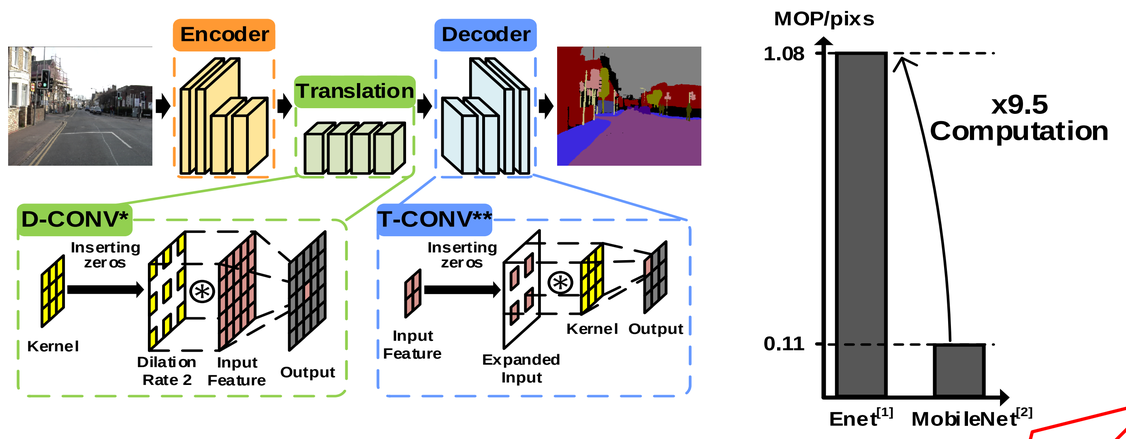
Application for Image Segmentation

<Autonomous Driving> <Action Recognition> <Human Robot Interaction>

- Generation of the useful image information in pixel-level
- Implementation by encoder-decoder neural network

Encoder-decoder Neural Network

<Structure of Encoder-decoder Network> <Unit Computation of Networks>



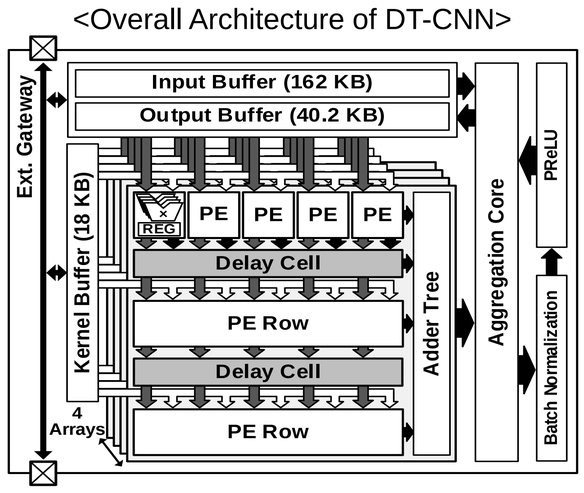
- Many virtual zeros in dilated convolution (D-CONV) and transposed convolution (T-CONV) → 86.6% of redundant virtual zero computations

[1] A. Paszke, et al., 2016 [2] A. Howard, et al., 2017

Zero-skip using the Delay Cell

Dilated and Transposed Convolution Accelerator (DT-CNN)

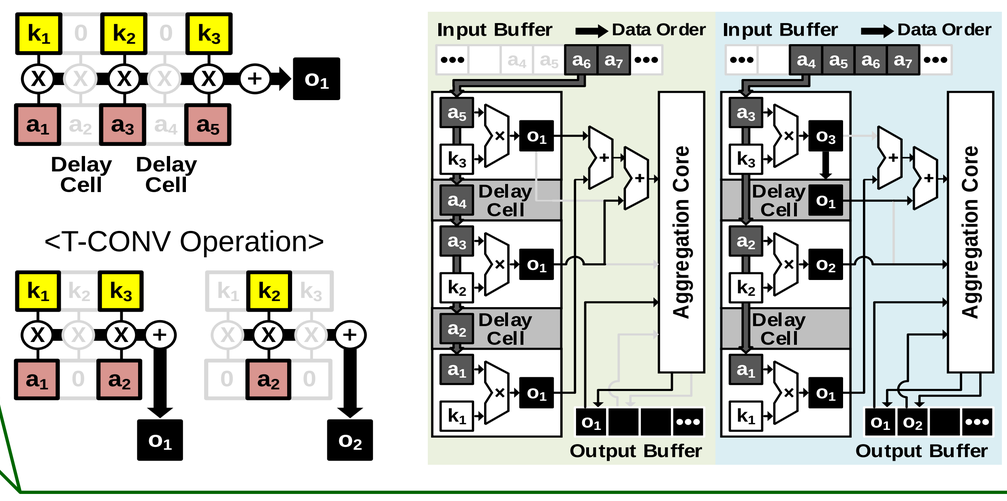
<Overall Architecture of DT-CNN>



- **Latching input features** in the delay cell for D-CONV
- **Latching output features** in the delay cell for T-CONV
- Performing zero-skip **with the simple delay cell logic**

Shared Data Path in D-CONV and T-CONV

<D-CONV Operation> <Data Path of D-CONV and T-CONV>

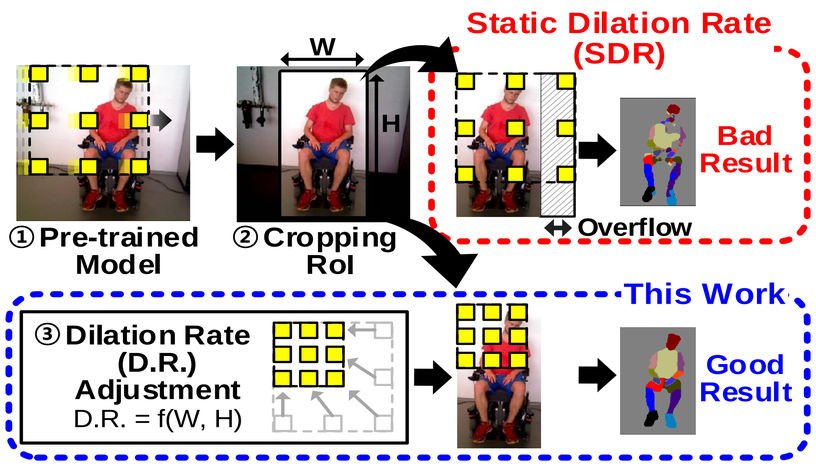


<T-CONV Operation>

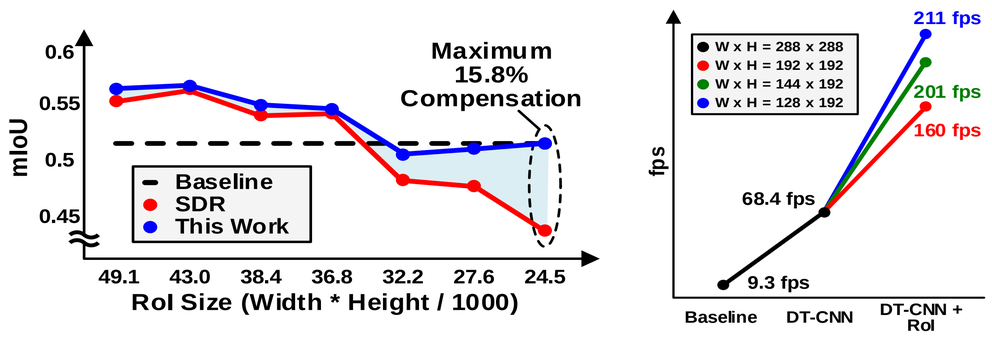
KAIST SSL

Dilation Rate Adjustment

ROI-based Segmentation with Dilation Rate Adjustment



Implementation Results of Dilation Rate Adjustment

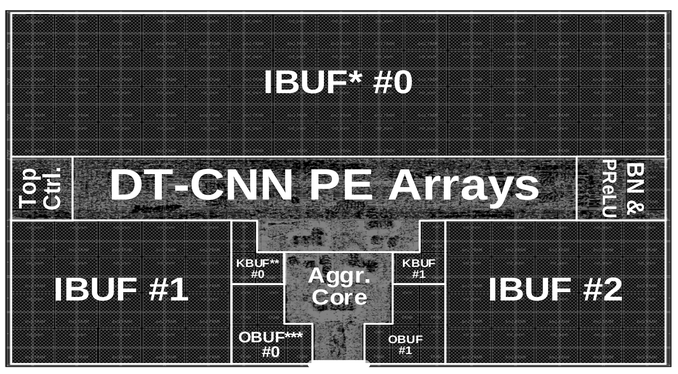


- Achieving **high throughput** by ROI-based image segmentation
- Compensation for the accuracy up to **15.8 percentage points**

[2] *Random feedback weights support learning in deep neural networks (Nature 2016)*

Implementation Results

Layout Photograph & Performance Summary

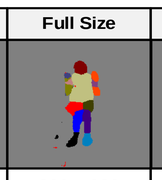
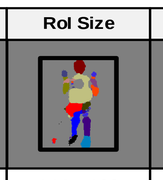
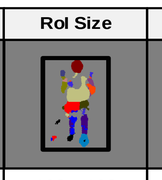


	This Work
Technology	65 nm
Supply Voltage	1.2 V
Clock Frequency	200 MHz
Throughput*	96 (8b)/ 639.7 (8b)*
Area	6.8 mm^2
Power	196 mW
Area Efficiency [GOPS/mm^2]	94.1*
Power Efficiency [TOPS/W]	3.26*
On-chip SRAM	220.5 KB

*Based on Logical Throughput of ENet

- **High throughput: 639.7 GOPS**
- **High energy-efficiency: 3.26 TOPS/W**
- **High area-efficiency: 94.1 GOPS/mm^2**

Visual Results of Image Segmentation*

	Test Image	Ground Truth	Full Size	RoI Size	This Work	RoI Size	This Work
Image							
Image Size (W x H)	288 x 288	288 x 288	288 x 288	144 x 192	144 x 192	128 x 192	128 x 192
Max. Dilation Rate	-	-	16	16	12	16	10
mIoU	-	-	0.5200	0.4659	0.5105	0.4345	0.5161
ΔmIoU	-	-	0%	- 10.4%	- 1.8%	- 16.4%	-0.7%
FPS	-	-	68.4	201		211	

- Achieving **211 frame-per-seconds** with high accuracy

*Freiburg sitting people dataset