HARD: Hardware-Aware lightweight Real-time semantic segmentation model Deployable from Edge to GPU

YoungWook Kwon[1*[0009−0009−0157−3023]], WanSoo Kim[1*[0009−0009−4176−6057]], and HyunJin Kim[1†[0000−0001−5017−3995]]

Dept. of Electronics and Electrical Engineering, Dankook University, 152, Jukjeon-ro, Suji-gu, Yongin-si, 16890, Gyeonggi-do, Republic of Korea
kyw96@naver.com, dhkstn115@naver.com, hyunjin2.kim@gmail.com

1 ImageNet-1k Pre-Train Setup

Table 1: Training hyper-parameters settings on ImageNet-1K.

config	value
optimizer	AdamW
initial learning rate	1e-6
weight decay	5e-3
optimizer momentum	β_1, β_2 = 0.9, 0.999
learning rate schedule	cosine decay
minimum learning rate	1e-6
warmup epochs	10
warmup learning rate	1e-3
training epochs	100
batch size	256
augmentation	RandAug(9, 0.5)
random resized crop	256
random flip	0.5

The encoder backbone of the proposed HARD is pre-trained on ImageNet-1K using two RTX 4090 GPUs. The hyper-parameters config settings shown in Table 1 are built according to those used in SOTA classification models.

2 Trade-Off with TITAN RTX

To demonstrate the high efficiency of HARD, we measured the inference speed of HARD and conventional methods using a single NVIDIA TITAN RTX. The results are shown in Table 2. The proposed HARD-GPU achieves high performance

*These authors contributed equally to this work.
†Corresponding Author

Table 2: Comparisons with other SOTA real-time methods on Cityscapes validation set. All FPS is measured on a single NVIDIA TITAN RTX GPU.

Resolution	Model	Params	FLOPs	mIoU	FPS	Resolution	Model	Params	FLOPs	mIoU	FPS
512×1024	ENet	0.38M	5.65G	58.3	58	512×1024	LiteHRNet	1.09M	4.66G	70.6	23
512×1024	FSSNet	0.29M	3.89G	58.8	117	512×1024	SeaFormer-S	4.1M	1.8G	70.7	73
512×1024	ESPNet	0.36M	4.1G	60.3	152	1024×2048	SGCPNet	0.61M	4.5G	70.9	78
512×1024	MiniNet	1.41M	6.71G	61.5	271	512×1024	FBSNet	0.61M	22.06G	70.9	16
360×640	**HARD-XXS**	0.11M	0.93G	64.1	**282**	512×1024	EdgeNet	-	-	71.0	31
1024×2048	CGNet	0.50M	28.0G	64.8	39	512×1024	FDDWNet	0.77M	12.38G	71.5	73
512×1024	NDNet	0.50M	3.9G	65.1	130	512×1024	MiniNetV2	0.51M	9.26G	71.8	98
512×1024	ESPNetV2	1.25M	5.65G	66.2	105	512×1024	MSCFNet	1.15M	17.1G	71.9	50
512×1024	EDANet	0.69M	8.88G	67.3	121	512×1024	STDC1	12.5M	23.1G	72.2	126
512×1024	ADSCNet	0.51M	12.68G	67.5	125	512×1024	SeaFormer-B	8.7M	3.1G	72.2	56
512×1024	ERFNet	2.06M	29.93G	68	76	360×640	LETNet	0.95M	13.59G	72.8	23
1024×2048	FastSCNN	1.14M	6.72G	68.6	146	512×1024	SCTNet-S	4.6M	7.1G	72.8	121
360×640	**HARD-XS**	0.39M	1.93G	69.6	**244**	360×640	**HARD-S**	0.76M	3.24G	72.8	**190**
1024×2048	CFPNet	0.27M	21.07G	70.1	30	512×1024	PP-LiteSeg-T	4.4M	4.3G	73.1	132
512×1024	FPENet	0.4M	12.8G	70.1	113	512×1024	BiseNetv2	5.2M	35.5G	73.4	124
512×1024	SwiftNetRN	12.1M	32.1G	70.2	101	512×1024	AFFormer-Base	3.0M	8.6G	73.5	50
512×1024	LEDNet	0.94M	11.31G	70.6	66	512×1024	**HARD-GPU**	3.9M	11.5G	**73.8**	**162**

and fast inference speed even though it is a 3-branch real-time semantic segmentation model. Compared to the state-of-the-art performance model, SCTNet, we achieved improved performance and 41 FPS faster inference speed. HARD-XXS achieved the fastest inference performance with 282FPS. The proposed HARD shows robust performance on any device.

3 Peak Memory Analysis on MCU

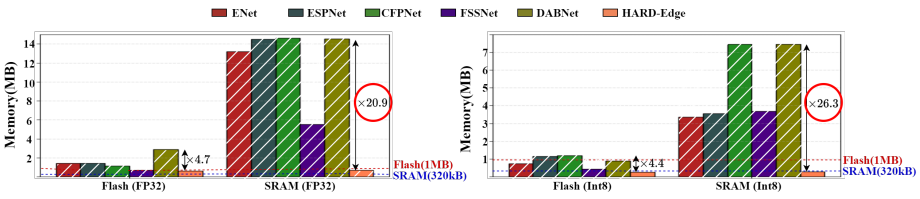

Fig. 1: Peak memory analysis between previous lightweight segmentation models.

HARD-Edge was deployed on the STM32F746NG, along with other lightweight segmentation models, to demonstrate memory efficiency on the MCU. Compared to DABNet on FP32, HARD-Edge shows ×20.9 less SRAM usage. When quantized with Int8, it uses ×26.3 less SRAM, at 283.05 kB, making it the only existing model that uses less SRAM than is available on the MCU.

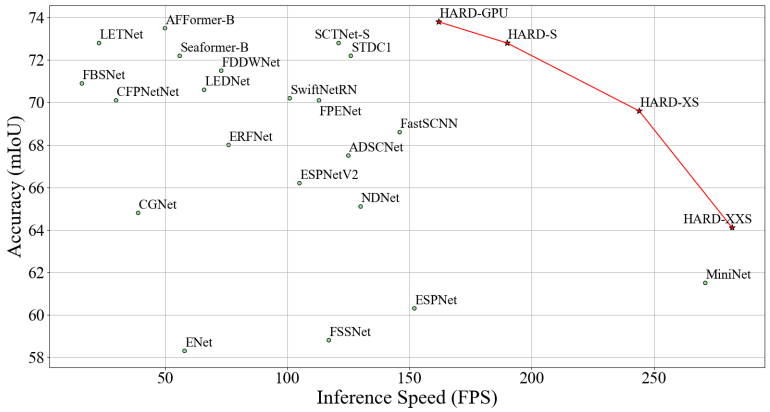

Fig. 2: Inference speed and accuracy performance on Cityscapes validation set using a single NVIDIA TITAN RTX.

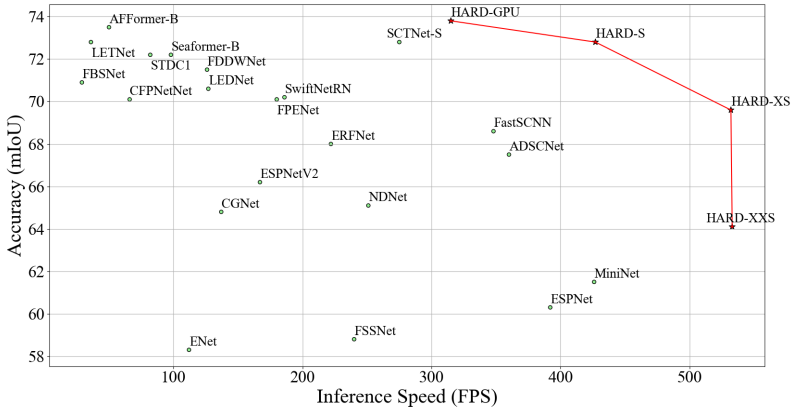

Fig. 3: Inference speed and accuracy performance on Cityscapes validation set using a single NVIDIA RTX 4090.