Efficient Parallel Multi-Scale Detail and Semantic Encoding Network for Lightweight Semantic Segmentation

Xiao Liu, Xiuya Shi, Lufei Chen, Linbo Qing and Chao Ren

College of Electronics and Information Engineering, Sichuan University, Chengdu, China

Codes Wechat

Motivations and Contribution

Goal: Learning the human brain uses the hierarchical organization of neurons to process visual information to achieve detailed local information and coarse large-range relationships extraction in parallel, enabling the recognition of object boundaries and object-level areas.

Introduction:

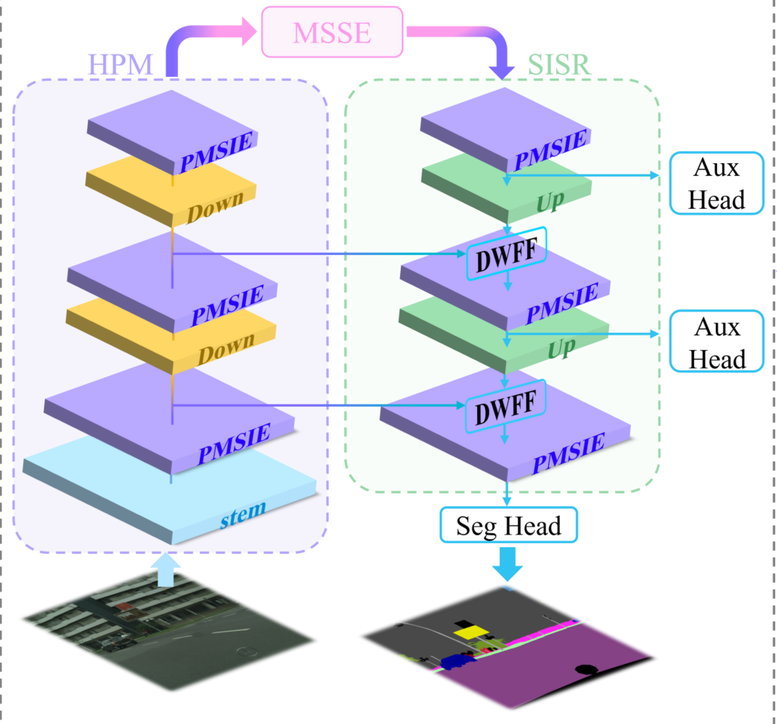
- How to integrate information from multiple scales and hierarchies as the human brain for efficient and robust visual perception and processing?
- Most excellent backbone networks use a hierarchical feature pyramid style, each layer of the network contains a relatively single scale of features, which is insufficient for dense prediction tasks that require rich multi-scale information for accurately identifying and segmenting objects in various contexts and scales.
- Most elaborated multi-scale semantic context modules are typically inserted at the end of the backbone network to extract multi-scale semantic contextual information. However, they are only inserted at the end of the pre-trained backbone network, which may not be sufficient to extract multi-scale spatial and semantic context in all cases.

Contributions:

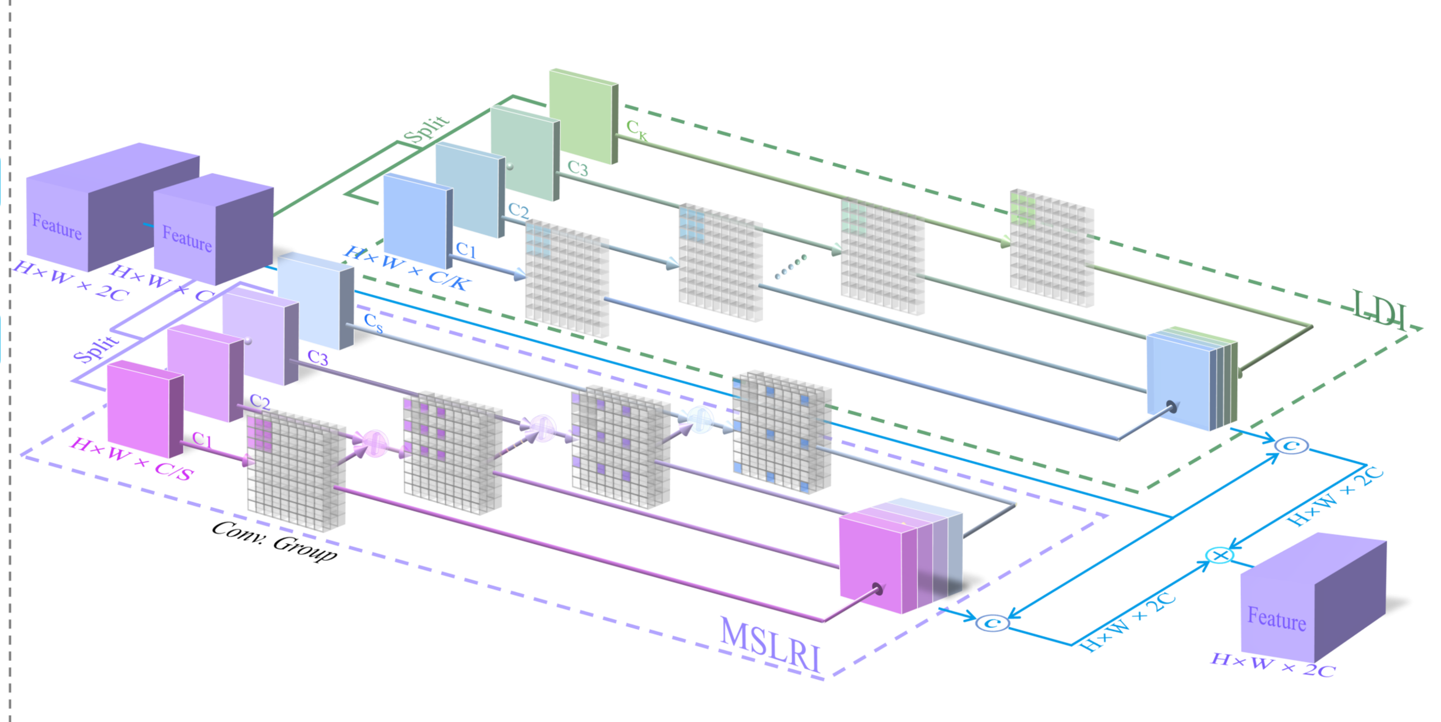
- This work focuses on developing more advanced multi-scale spatial and semantic context modules and improving the integration of these modules with hand-crafted multi-scale backbone networks to achieve more effective and efficient feature extraction for semantic segmentation.
- The proposed PMSDSEN incorporates PMSDSE and MSSE in the encoder-decoder architecture, effectively realizing from low-level feature extraction to high-level semantic interpretation at different scales and in different contexts.
- The experiment results present that PMSDSEN can achieve better results on various segmentation benchmarks, including Cityscapes and CamVid. For example, PMSDSEN achieves 73.2% mIoU on the Cityscapes test set with 0.9M parameters.

Method

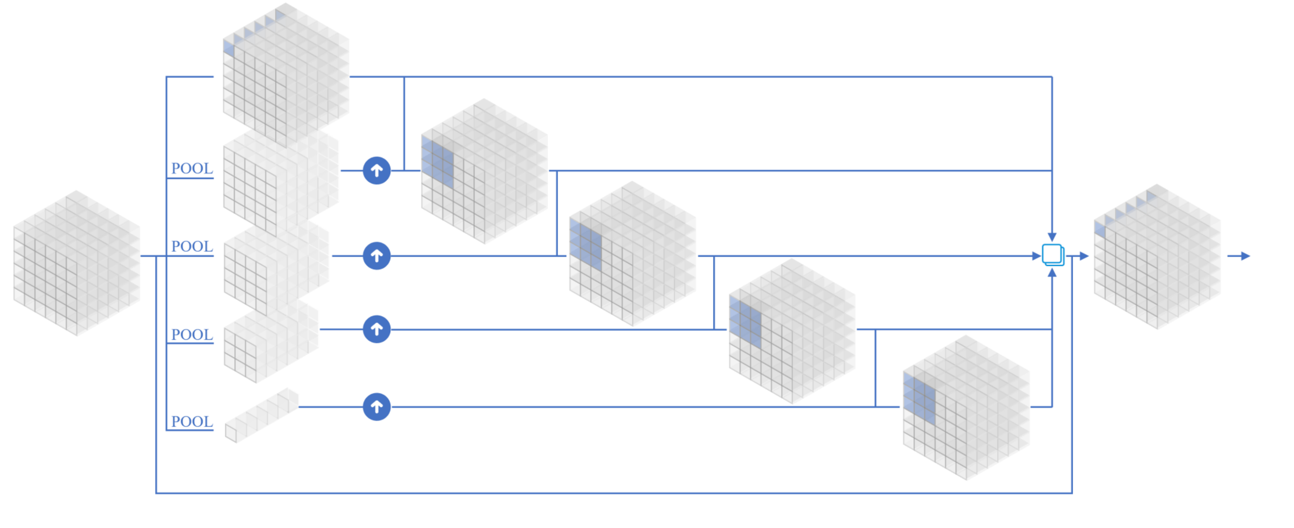
Network Architecture:



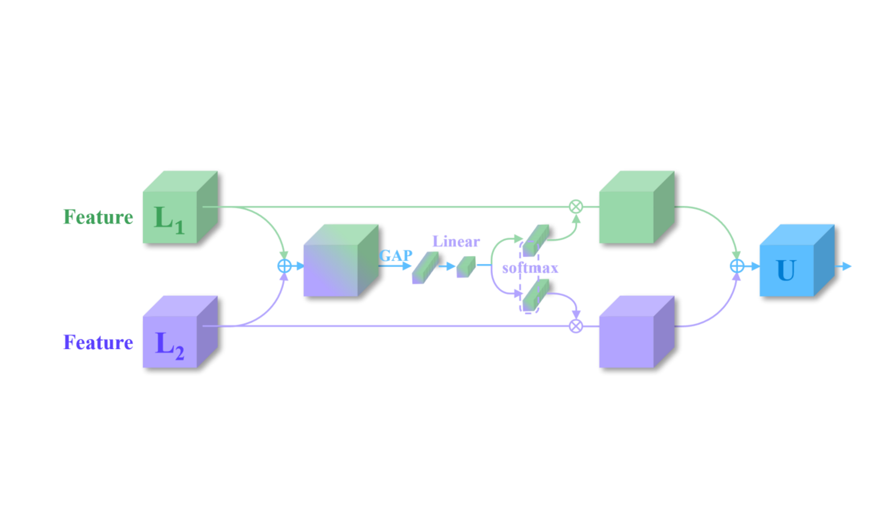
(a) PMSDSEN



(b) PMSDSE



(c) MSSE



(d) DWFF

Conclusion

- Achieving a more optimal trade-off between model complexity and performance.
- From rich and detailed local information to coarse and complex large-range relationships, from fine-grained details and textures to abstract category and semantic information, network achieve efficiently low-level feature extraction and high-level semantic interpretation at different scales and in different contexts.

Contact

Hey! I'm Xiao Liu, a Master at Sichuan University. My research interests include semantic segmentation and image restoration. Should you have any question, please contact at liuxmail1220@gmail.com.

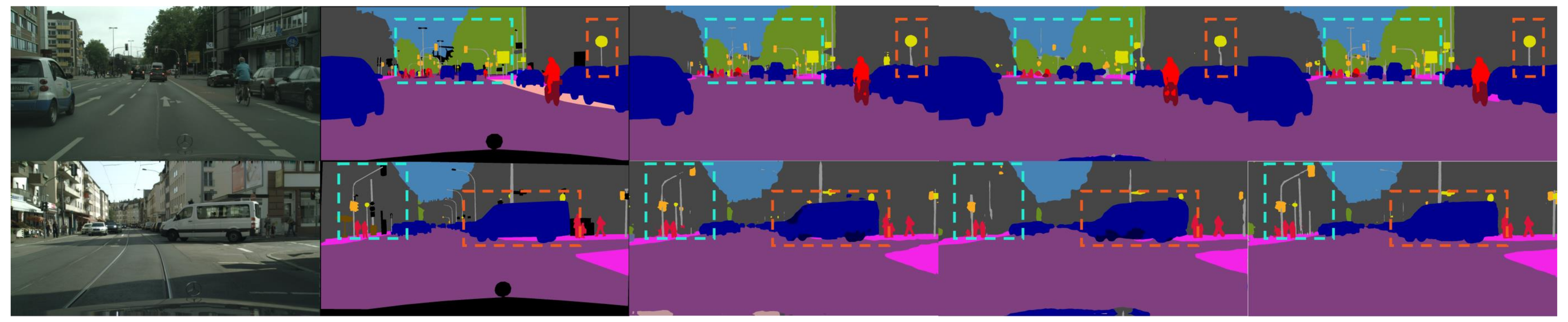
Results

Quantitative comparisons on the Cityscapes dataset:

Table 2: Comparisons with other state-of-the-art methods on Cityscapes dataset.

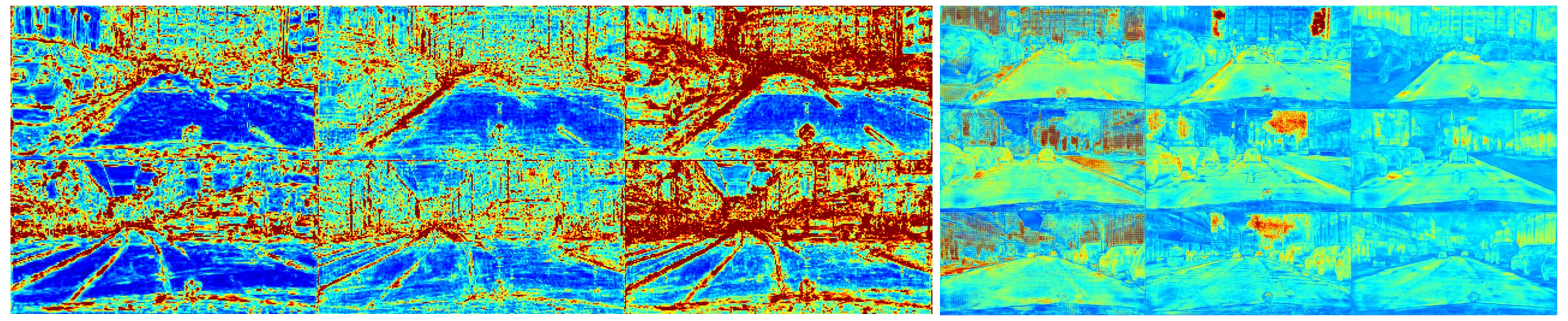
Methods	Year	Resolution	Backbone	Params(M)	Multi-Adds(G)	FPS	mIoU(%) val	mIoU(%) test
DeepLab [2]	2015	512×1024	ResNet-101	262.10	457.8	0.25	-	63.5
SQNet [29]	2016	1024×2048	SqueezeNet	-	270.0	17	-	59.8
ENet [24]	2016	512x1024	no	0.36	3.8	76	-	58.3
SegNet [1]	2017	640×360	VGG-16	29.50	286.0	17	-	57.0
ERFNet [28]	2017	512×1024	no	2.10	-	42	70.0	68.0
ICNet [38]	2018	1024×2048	PSPNet50	26.5	28.3	30	67.7	69.5
ESPNet [21]	2018	512×1024	ESPNet	0.36	-	113	-	60.3
BiSeNet-v1 [36]	2018	768×1536	Xception39	5.80	14.8	106	69.0	68.4
ESPNet-v2 [22]	2019	512×1024	ESPNet-v2	-	2.7	80	66.4	66.2
LEDNet [30]	2019	512×1024	no	0.94	-	40	-	70.6
FPENet [19]	2019	512×1024	no	0.40	12.8	55	-	70.1
DABNet [16]	2019	1024×2048	no	0.76	10.5	28	-	70.1
CAS [37]	2019	768×1536	no	-	-	108	71.6	70.5
CGNet [31]	2020	360×640	no	0.50	6.0	-	-	64.8
NDNet [34]	2021	1024×2048	no	0.50	14.0	40	-	65.3
CFPNet [20]	2021	1024×2048	no	0.55	-	30	-	70.1
EdgeNet [11]	2021	512×1024	no	-	-	31	-	71.0
MGSeg [13]	2021	1024×1024	ShuffleNet-v2	4.50	16.2	101	-	72.7
BiSeNet-v2 [35]	2021	512×1024	no	3.40	21.2	156	73.4	72.6
STDC1-Seg50 [5]	2021	512×1024	STDC1	8.40	23.19	250	72.2	71.9
FPANet [32]	2022	512×1024	no	14.10	-	-	-	72.0
SGCPNet [12]	2022	1024×2048	no	0.61	4.5	103	-	70.9
FBSNet [8]	2022	512×1024	no	0.62	9.7	90	-	70.9
MSCFNet [9]	2022	512×1024	no	1.15	17.1	50	-	71.9
PMSIEN (Ours)	2023	512×1024	no	0.92	10.2	53	73.6	73.2

Qualitative comparisons on the Cityscapes dataset:



Input Image Ground Truth BiSeNetv2 STDC1 PMSDSEN(Ours)

Visual analysis:



(a) LDI (b) MSLRI (c) Fusion (a) DWFF (b) Concat (c) Sum

Visualization of features for each branch in the PMSDSE (Left), and different fusion strategies (Right). PMSDSE can extract rich and detailed local information, as well as coarse and complex large-range relationships parallelly. Therefore, the fusion features possess finely detailed localization and powerful long-range relationships. DWFF enables network to focus on the most informative parts of feature map by comparing the darker parts of the feature map.