Ultra High Resolution in CVPR2022

Speaker: Gong, Qiqi

Outline

- Introduction to Ultra High Resolution
- Method Summary
- Paper Sharing
- Inspiration

Introduction to UHR (Ultra high resolution,超高分辨率)

- Semantic Segmentation Dataset Comparison

Name	Train #	Val #	Test #	Class #	Resolution
Pascal VOC	1,464	1,449	-	21	<1000
COCO	118K	5K	41K	91	640 X 480
ADE20K	20K	2K	3K	150	<1024 x 768
Cityscapes	2975	500	1525	20	1024 X 2048
Mapillary Vistas	18K	2K	5K	124	1024*768~4000*6000
BIG	-	-	-	-	2048×1600~5000×3600
UHRSD	4932	988	-	-	4K~8K

Method Summary

- Background: Development of collecting devices brings UHR images
- Fundamental Problems
  - Two obstacles: Computation & Memory
  - Receptive field
  - Downsampling

Method Summary

- Multi-scale Decoder
  - GLNet (CVPR2019)
  - CascadePSP (CVPR2020)
- Boundary Refinement
  - BASNet (CVPR2019)
  - DeepStrip (CVPR2020)
  - SegFix (ECCV2020)
  - PointRend (ECCV2020)
- Bi-path
  - PGNet (CVPR2022)
  - ISDNet (CVPR2022)

Paper Sharing——PGNet

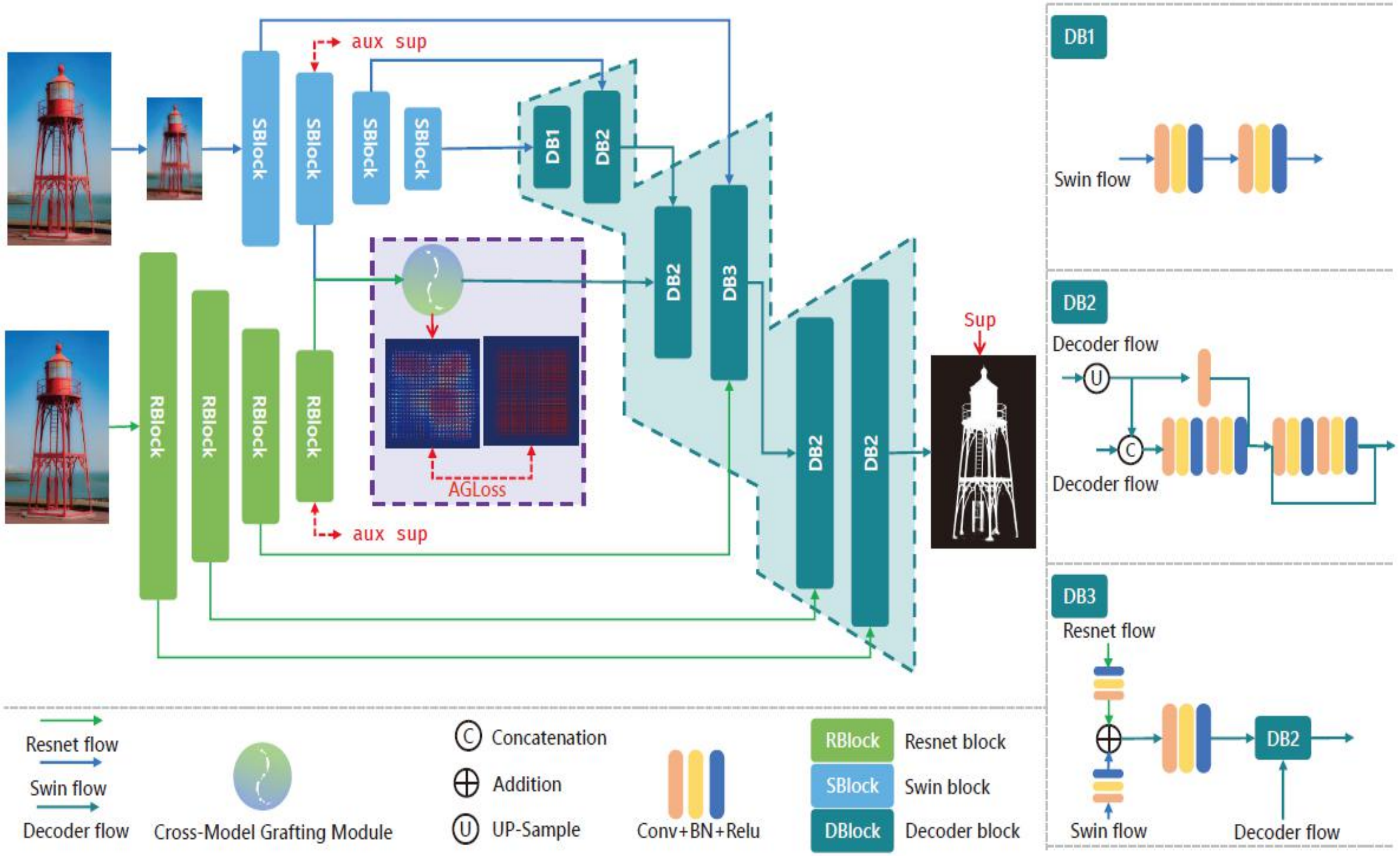
Pyramid Grafting Network for One-Stage High Resolution Saliency Detection

Chenxi Xie[1], Changqun Xia[*2], Mingcan Ma[1], Zhirui Zhao[1], Xiaowu Chen[1,2], Jia Li[1,2]
[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2]Peng Cheng Laboratory, Shenzhen, China
{xiechenxi,mingcanma,zhiruizhao,chen,jiali}@buaa.edu.cn, xiachq@pcl.ac.cn

PGNet

- Motivation
  - Transformer encoder performs well in LR cases
  - CNN encoder performs well in HR cases
- Key point: How to fuse information from different encoders?

PGNet



aux sup
SBlock
SBlock
SBlock
SBlock
DB1
DB2
DB2
DB3
DB2
DB2
Sup
RBlock
RBlock
RBlock
RBlock
AGLoss
aux sup
DB1
Swin flow
DB2
Decoder flow
U
C
Decoder flow
DB3
Resnet flow
DB2
Swin flow
Decoder flow
Resnet flow
Swin flow
Decoder flow
Cross-Model Grafting Module
C Concatenation
Addition
U UP-Sample
Conv+BN+Relu
RBlock Resnet block
SBlock Swin block
DBlock Decoder block

PGNet

- Cross Model Grafting (融合) Module

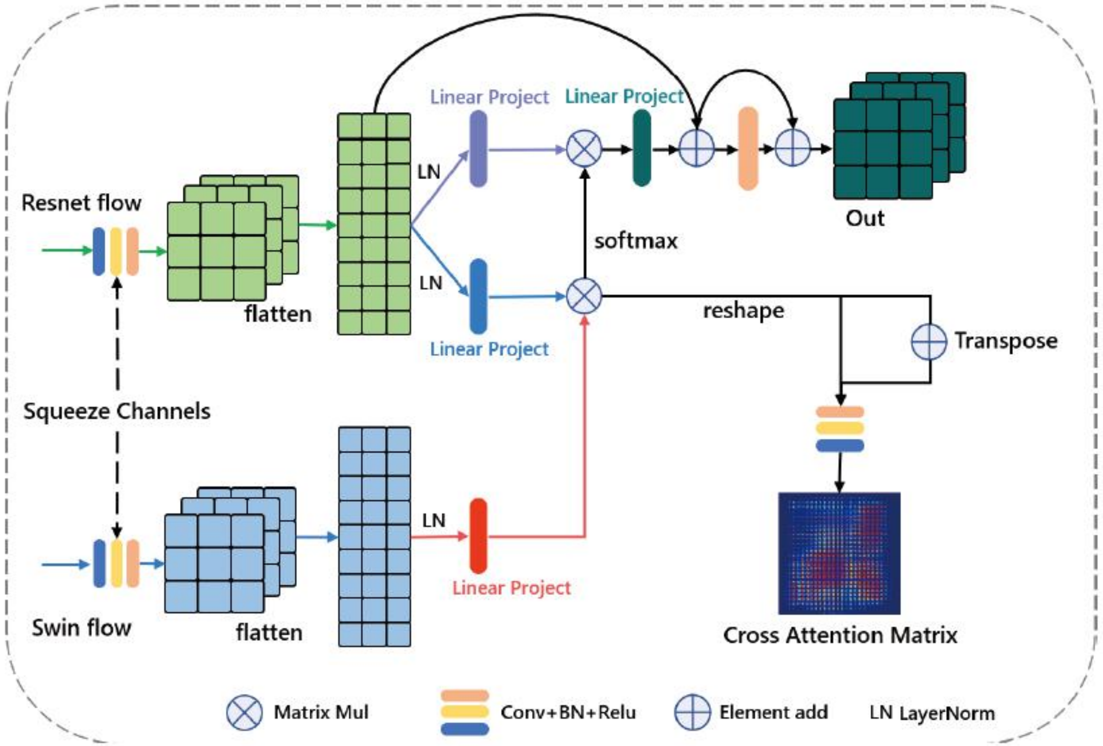
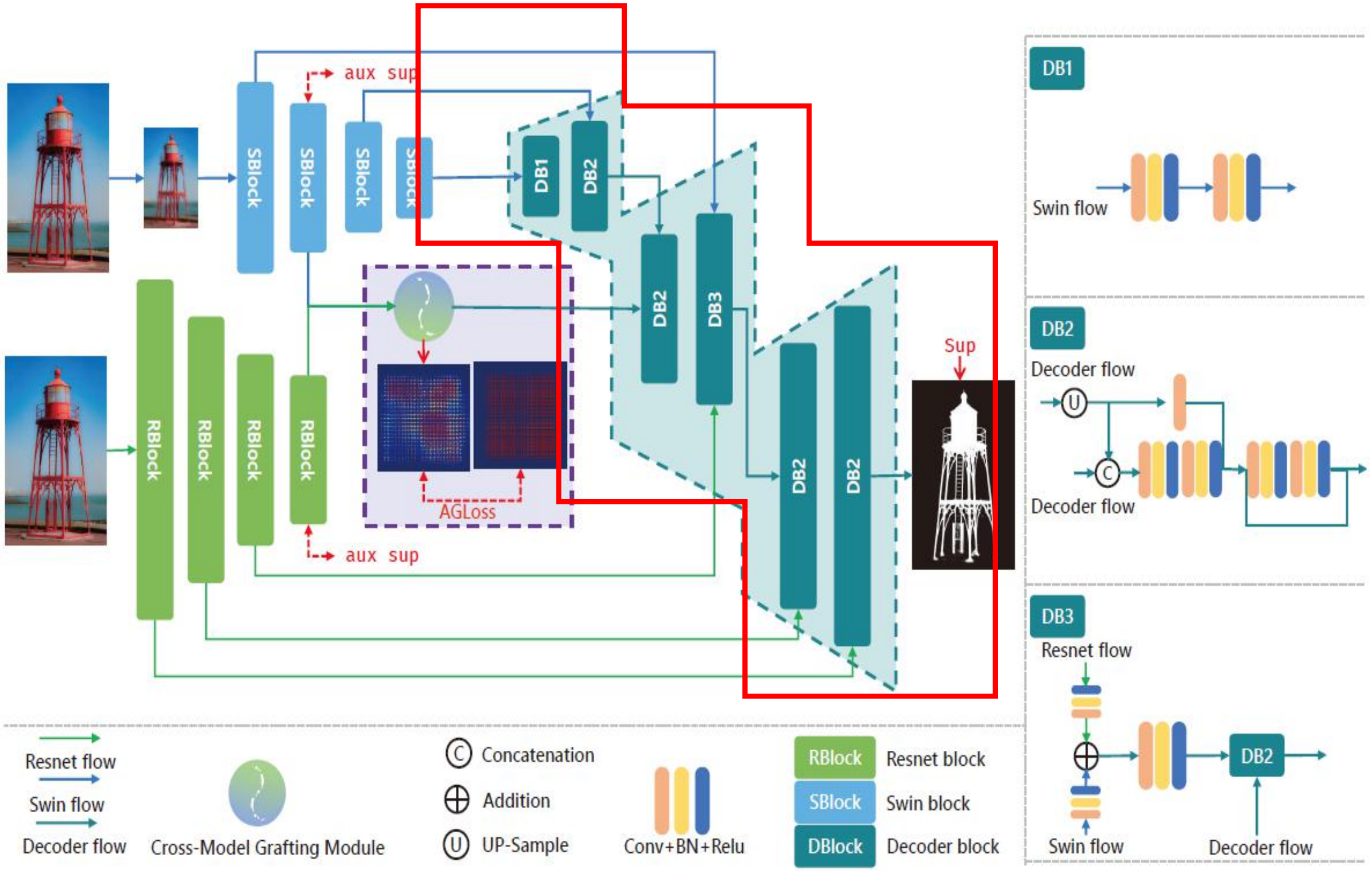


Figure 5. Architecture of Cross-Model Grafting Module.

PGNet

- Decoder



ISDNet

ISDNet: Integrating Shallow and Deep Networks for Efficient Ultra-high Resolution Segmentation

Shaohua Guo[1*], Liang Liu[2*], Zhenye Gan[2], Yabiao Wang[2], Wuhao Zhang[2],
Chengjie Wang[2], Guannan Jiang [5], Wei Zhang[5], Ran Yi[1†], Lizhuang Ma[1,3†], Ke Xu [4]
[1]Shanghai Jiao Tong University [2]Youtu Lab, Tencent
[3]East China Normal University [4]City University of Hong Kong [5]CATL
{guoshaohua, ranyi}@sjtu.edu.cn;{jianggn,zhangwei}@catl.com; ma-lz@cs.sjtu.edu.cn;
{leoneliu, winggzygan,caseywang, wuhaozhang,jasoncjwang}@tencent.com; kkangwing@gmail.com;

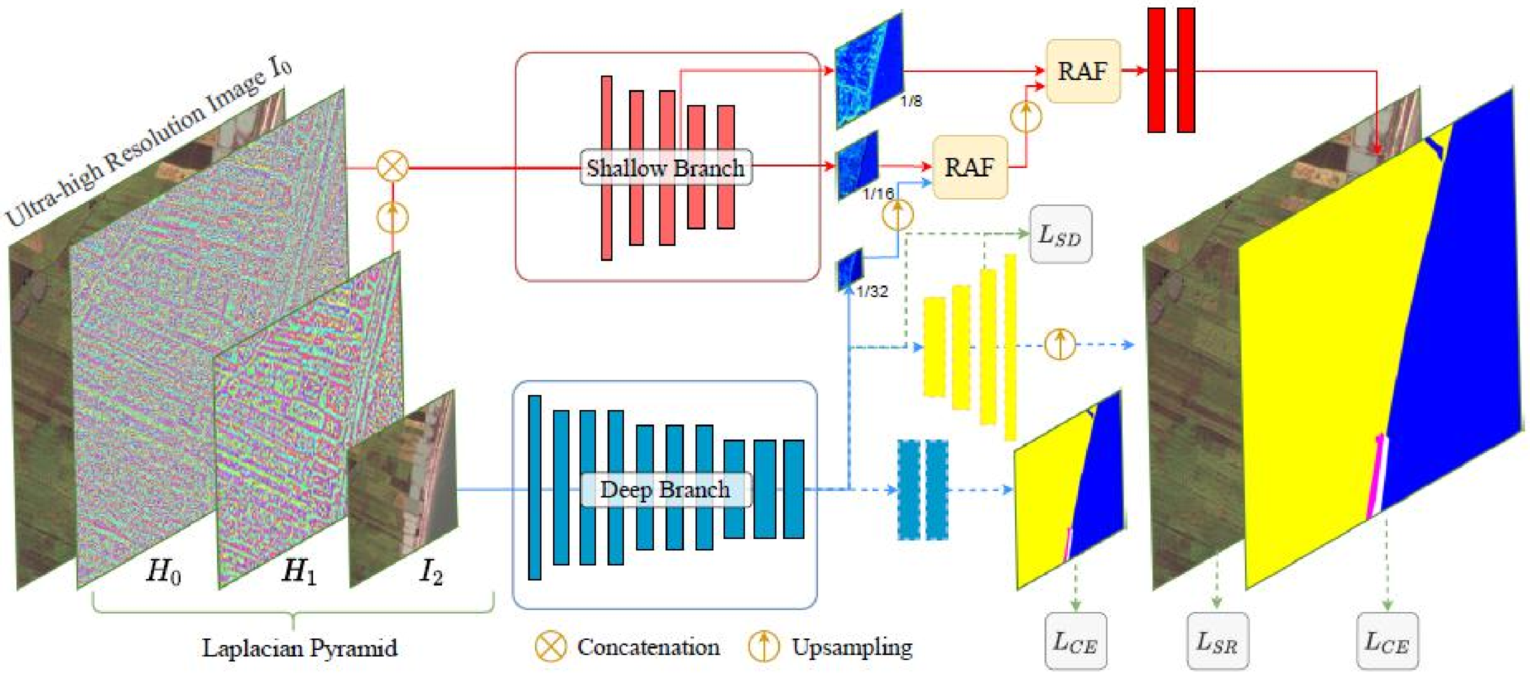
ISDNet

- Summary
  - Bilateral path (shallow + deep)
  - Relation-Aware feature Fusion mechanism (RAF)
  - **Super-resolution aux loss**
- Motivation
  - Trade-off among accuracy, speed and memory for UHR sem_seg

Method	mIoU ↑	FPS ↑	Memory(MB) ↓
Generic Methods			
BiSeNetV1 [37]	74.44	42.43	2147
BiSeNetV2 [36]	75.80	43.07	1602
PSPNet [40]	74.87	15.15	1584
ICNet [39]	74.43	**68.55**	**1390**
STDC [11]	74.5	62.15	1536
DeepLabv3 [1]	**76.70**	13.32	1468
UHR Methods			
DenseCRF [21]	62.95	0.04	1575
DGF [32]	63.33	3.13	1727
SegFix [38]	65.83	2.63	2033
PointRend [20]	64.39	7.14	2052
MagNet [18]	67.57	0.34	2007
MagNet-Fast [18]	66.91	3.13	2007
Ours (ISDNet)	**76.02**	**50.79**	**1510**

Table 3. Segmentation results on the CityScapes dataset. We evaluate the speed and memory under our environment, and the accuracy of UHR competitors are collected from [18].

ISDNet

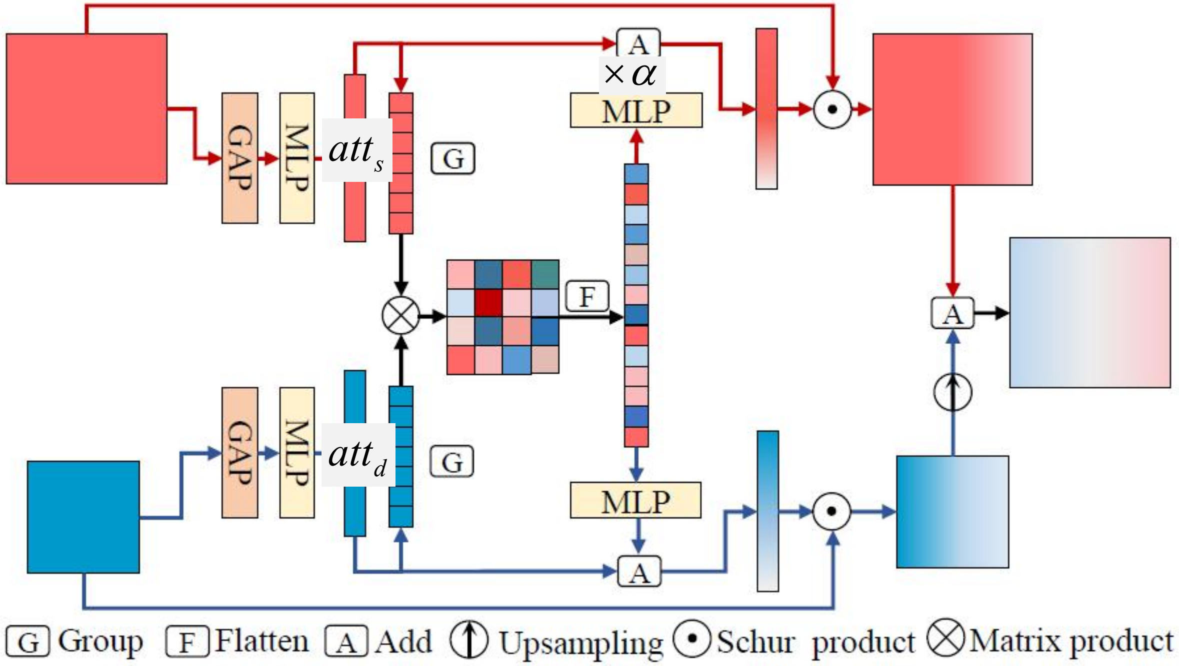


$$H_i = g_i(I) - \text{Upsample}\left(g_{i+1}(I)\right), \qquad (1)$$

where I represents the full scale image, $g(.)$ denotes guassian blur and i is the number of levels in the pyramid.

*STDC: 《Rethinking BiSeNet for Real-time Segmentation》--2021CVPR

ISDNet--RAF



ADD	CAT	CW	M_s	M_d	mIoU	FPS	Mem(MB)
✓			-	-	72.20	31.69	-
✓		✓	-	-	72.42	29.73	1891
	✓		-	-	71.88	23.98	-
	✓	✓	-	-	72.57	25.76	2204
✓		✓	✓		72.63	28.93	-
✓		✓	✓	✓	73.30	27.70	1948

Table 6. Comparison of feature fusion methods. ADD and CAT represent two simple fusion strategies: addition and concatenation. CW means channel-wise attention mechanism. M_s and M_d denote the relation-aware attention for deep and shallow branch.

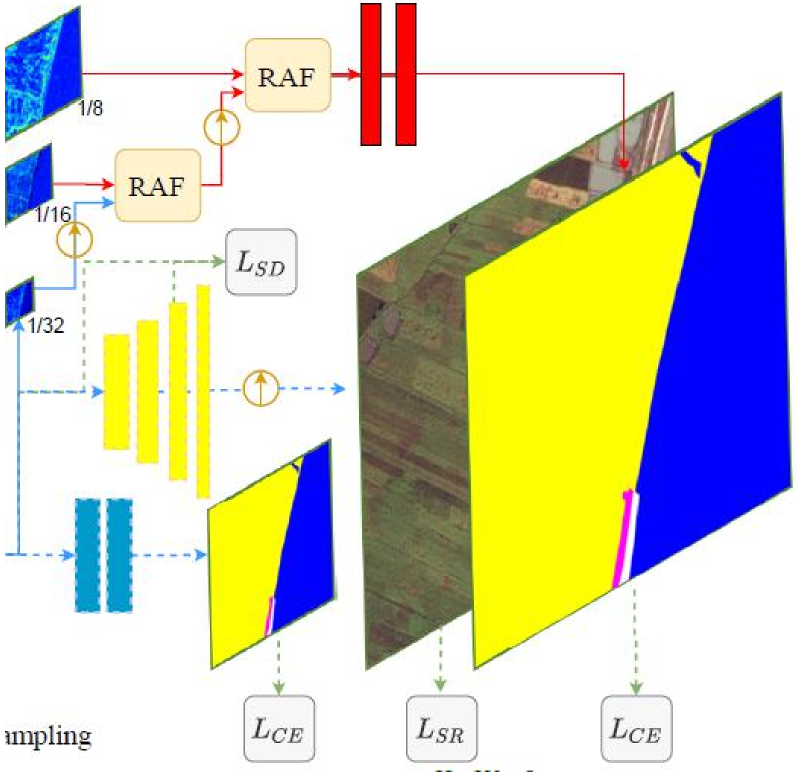
ISDNet--Loss

- Final Segmentation loss: L_{SEG}
- Deep branch aux loss: L_{AUX}
- Super-resolution loss: $L_{SR} = \left\| I_0 - I_{rec} \right\|_2^2$
- Structure distillatoin loss:

$$\mathcal{L}_{SD} = \left\| F_d^T F_d - F_{sr}^T F_{sr} \right\| .$$

Baseline	$\mathcal{L}_{SR}$	$\mathcal{L}_{SD}$	H	mIoU
✓				72.31
✓	✓			72.55
✓	✓	✓		72.70
✓	✓	✓	✓	73.30

Table 7. Comparison of loss components and heterogeneous input. H indicates high-frequency residual inputs for the shallow branch.



ISDNet--Ablation on Cityscapes

Method	mIoU	FPS	Mem(MB)
PSPNet [40]	74.87	15.15	1584
PSPNet [40] (½ scale)	72.87	54.99	1160
PSPNet [40] (¼ scale)	65.20	169.91	1076
PSPNet [40] + ISD	74.30	58.29	1540
Segformer-b0 [35]	73.45	13.70	3114
Segformer-b0 [35] (½ scale)	71.20	65.49	1174
Segformer-b0 [35] (¼ scale)	51.19	76.22	1032
Segformer-b0 [35] + ISD	72.99	41.82	1500

Table 4. Comparison of existing models integrating with our framework. We evaluate the corresponding methods with different scales to compare the accuracy and inference cost.

Inspiration

- CNN + Transformer structure
- Bi-path structure