01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

Yuval Nirkin*
Facebook AI & Bar-Ilan University

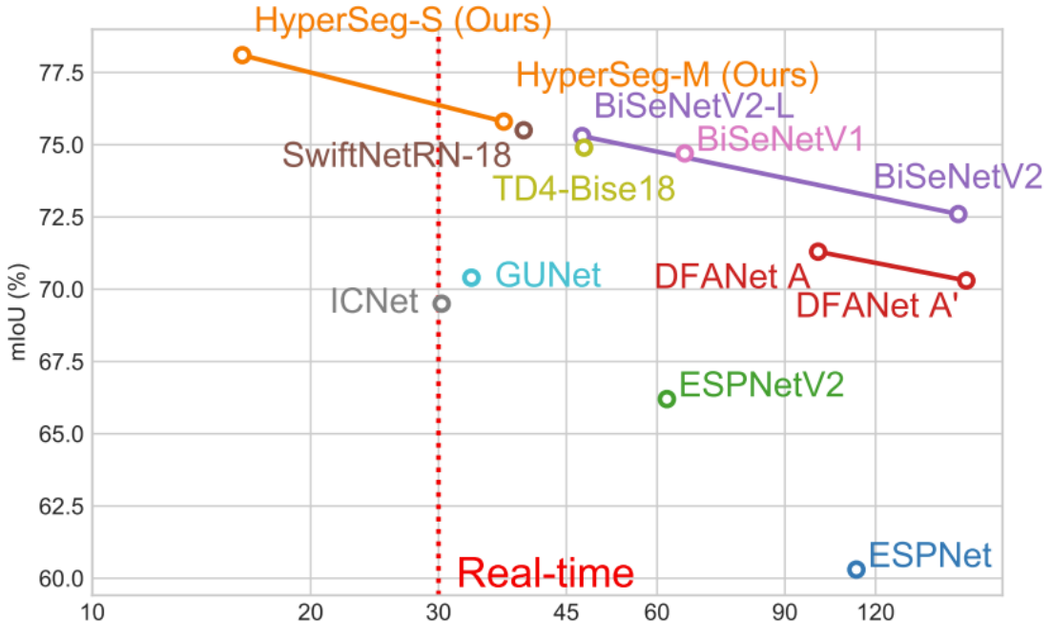
Lior Wolf
Facebook AI & Tel Aviv University

Tal Hassner
Facebook AI

- **CVPR 2021**

- **Contribution**

Scene understanding plays a crucial role in semantic segmentation,
Providing the network with additional adaptivity by hypernet.

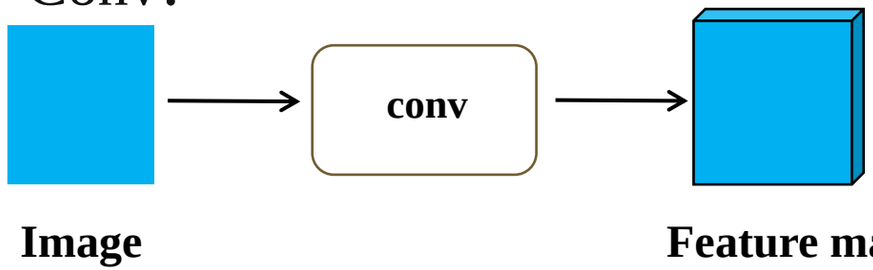


01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Meta-learning technique(**Referred to as dynamic networks or hypernet)

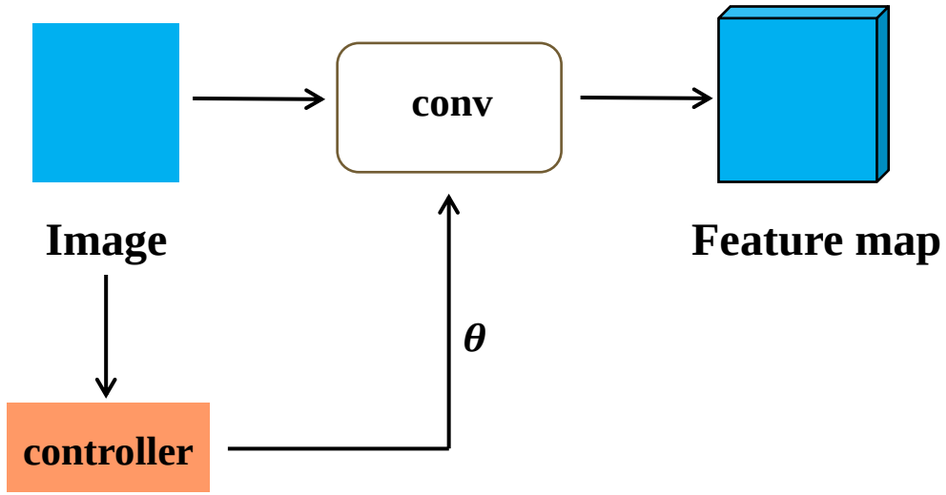
  - learn ‘how to learn a new knowledge’？（学会学习）

  - Conv：



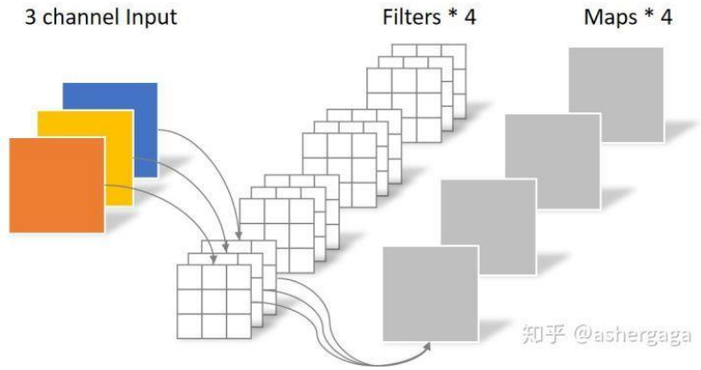

```
nn.Conv2d(in_channels, out_channels, kernel_size)
```

  - Hyper Conv:



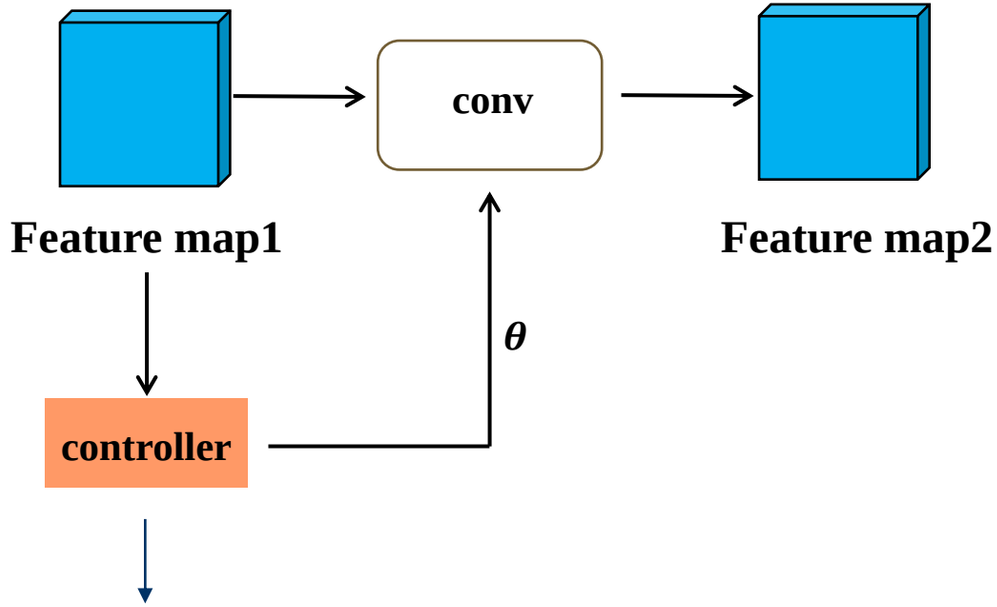

```
F.conv2d(x, w, b, stride=, dilation=, padding=,groups=)

X → feature map(batchsize, channel, h, w)
w → weight of conv(out_channel, in_channel, kh, kw)
b → bias of conv(out_channel)
```



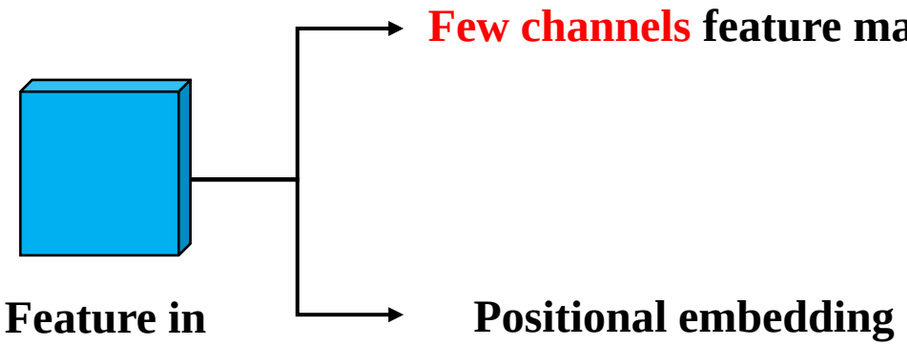
01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Meta-learning technique(**Referred to as dynamic networks or hypernet)
  - Hyper Conv:



(Shape) Batchsize×**Channel** × 1 × 1

Channel = out_channel × in_channel × k_h × k_w



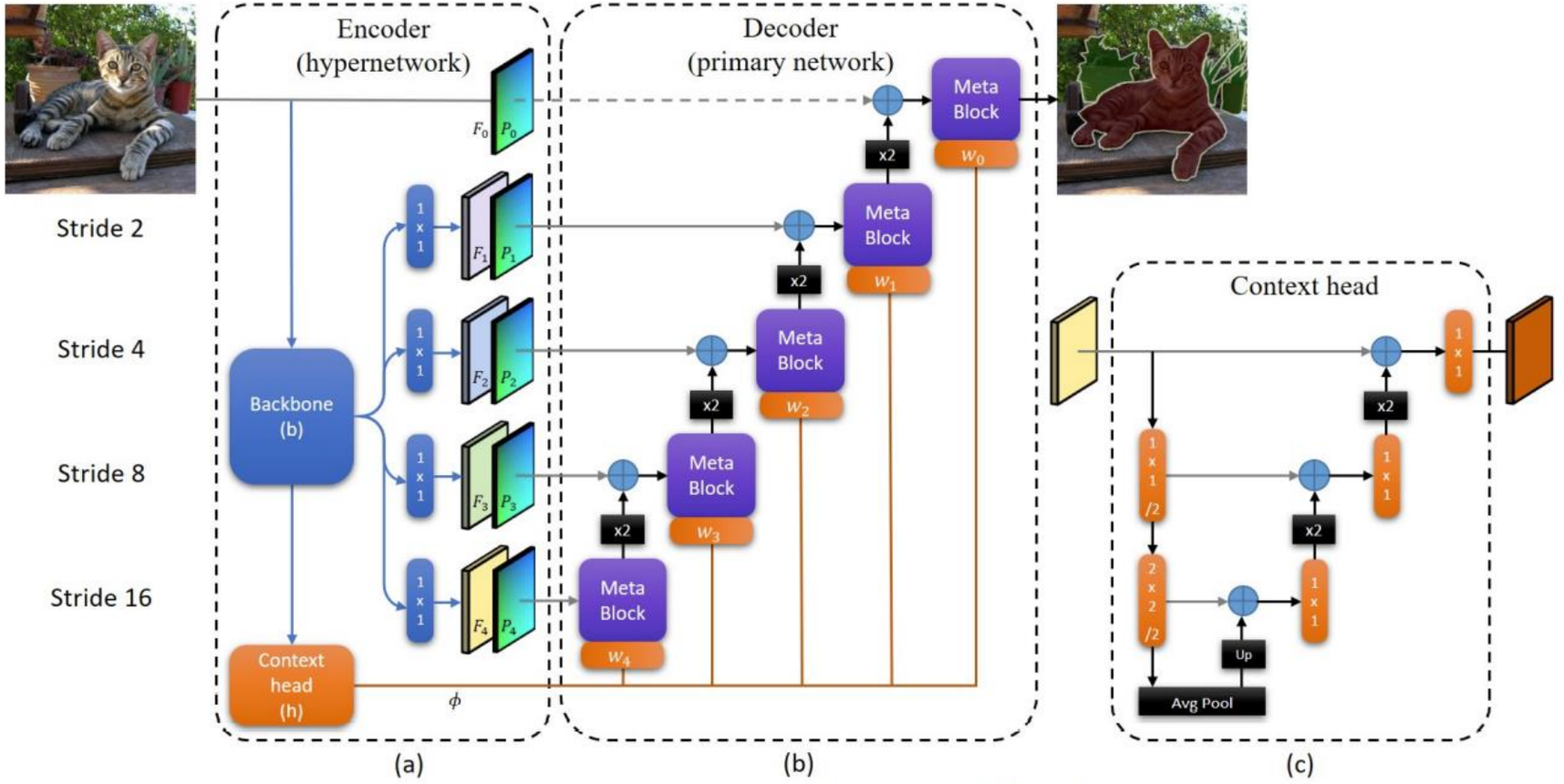
Positional embedding:(5x5)

tensor([[[[-1.0000, -0.5000, 0.0000, 0.5000, 1.0000],
[-1.0000, -0.5000, 0.0000, 0.5000, 1.0000],
[-1.0000, -0.5000, 0.0000, 0.5000, 1.0000],
[-1.0000, -0.5000, 0.0000, 0.5000, 1.0000],
[-1.0000, -0.5000, 0.0000, 0.5000, 1.0000]],

[[-1.0000, -1.0000, -1.0000, -1.0000, -1.0000],
[-0.5000, -0.5000, -0.5000, -0.5000, -0.5000],
[0.0000, 0.0000, 0.0000, 0.0000, 0.0000],
[0.5000, 0.5000, 0.5000, 0.5000, 0.5000],
[1.0000, 1.0000, 1.0000, 1.0000, 1.0000]]])

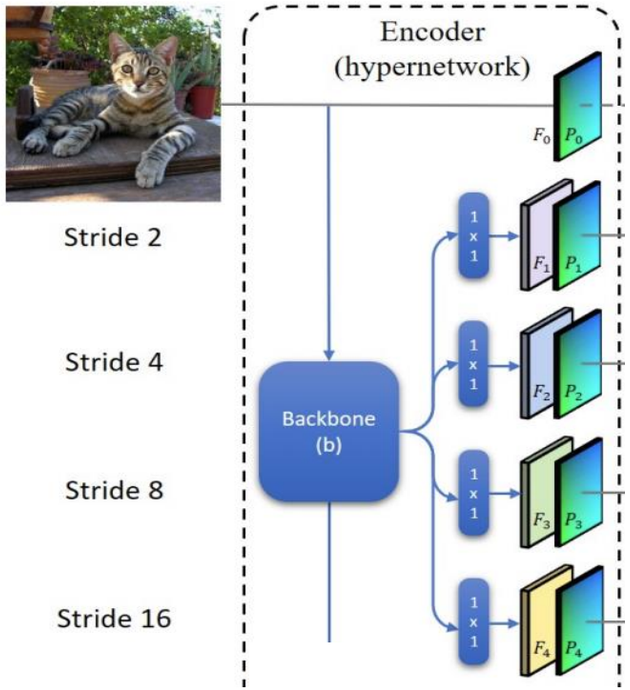
01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Overall network architecture**
  - Encoder(hypernetwork): EfficientNet/ ResNet18/ PSPNet50
  - Context head(weight mapper)
  - Decoder(meta block)



01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Backbone（EfficientNet / ResNet 18 / PSPNet50）**
  - Stride1 ->RGB map
  - Stride2、Stride4、Stride8、Stride16 -> 1×1conv -> feature map
  - Stride32 -> weight
  - Position embedding $P_{i,j}^{H,W} = (\frac{2i-H+1}{H-1}, \frac{2j-W+1}{W-1}), i \in [0, H), j \in [0, W)$, $P^{H,W} \in \mathbb{R}^{2\times H\times W}$



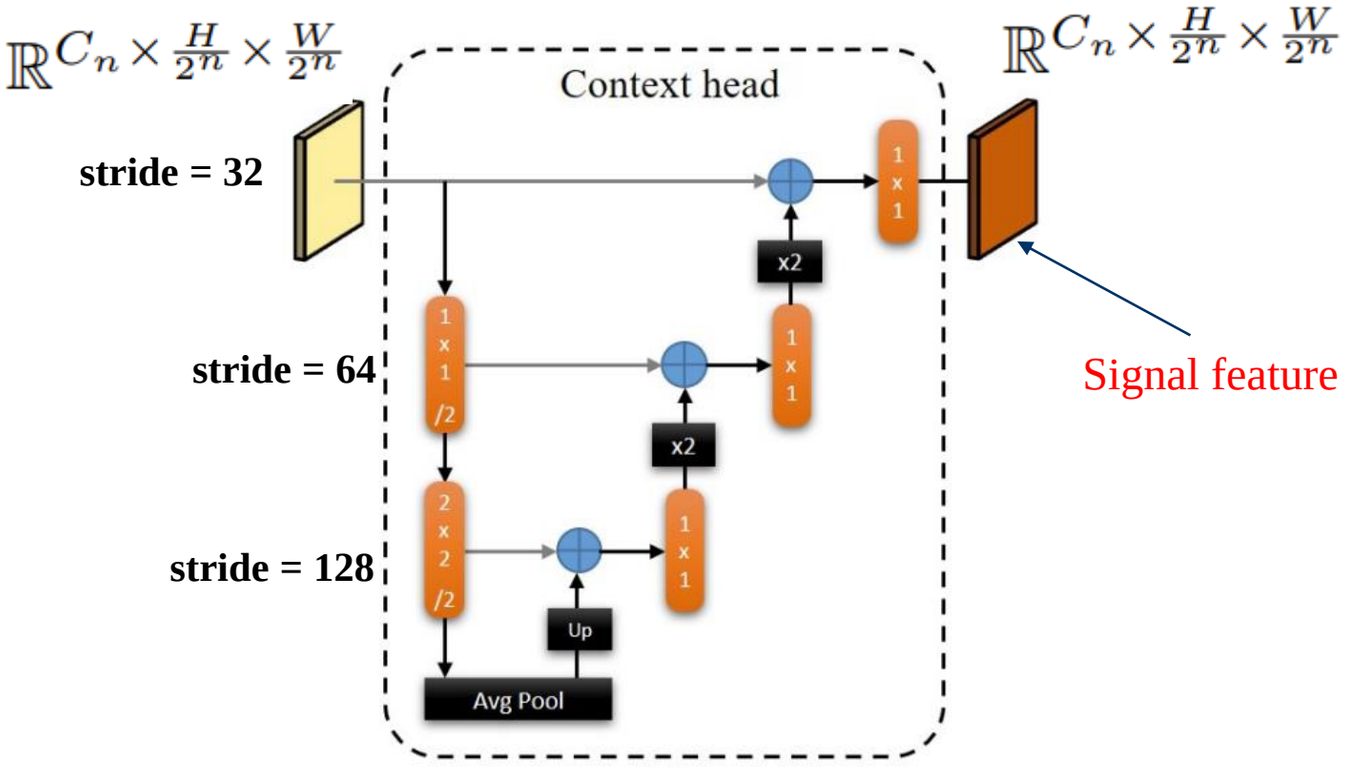
Positional embedding:(5x5)

tensor([[[[-1.0000, -0.5000, 0.0000, 0.5000, 1.0000],
[-1.0000, -0.5000, 0.0000, 0.5000, 1.0000],
[-1.0000, -0.5000, 0.0000, 0.5000, 1.0000],
[-1.0000, -0.5000, 0.0000, 0.5000, 1.0000],
[-1.0000, -0.5000, 0.0000, 0.5000, 1.0000]],

[[-1.0000, -1.0000, -1.0000, -1.0000, -1.0000],
[-0.5000, -0.5000, -0.5000, -0.5000, -0.5000],
[0.0000, 0.0000, 0.0000, 0.0000, 0.0000],
[0.5000, 0.5000, 0.5000, 0.5000, 0.5000],
[1.0000, 1.0000, 1.0000, 1.0000, 1.0000]]])

01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Context head**(stride = 32)
  - **Employ 2×2 convolutions with a stride of 2**
  - Computationally cheaper than 3 × 3 convolutions
  - Padding for the low-resolution feature maps significantly increase the spatial resolution



01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

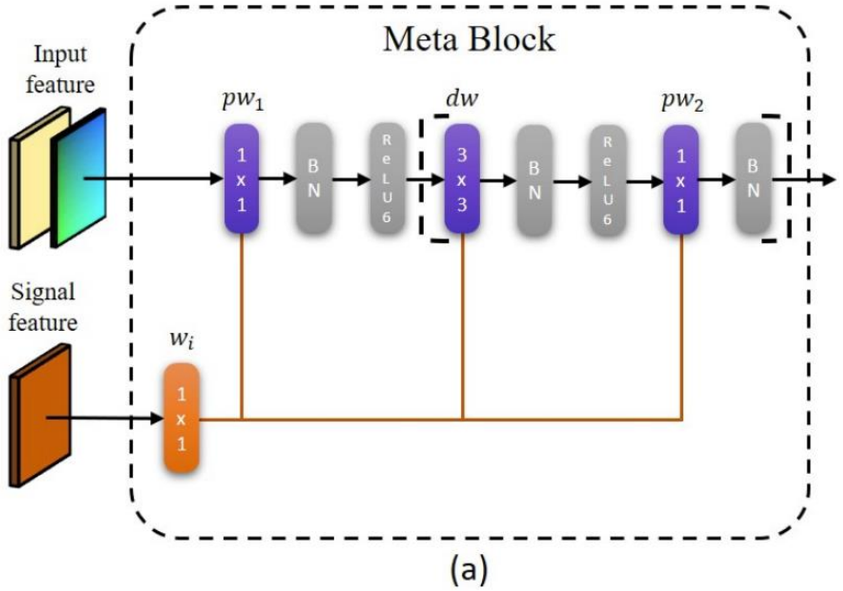
- **Meta block**
  - point-wise convolution(pw) + depth-wise convolution(dw) + point-wise convolution(pw)

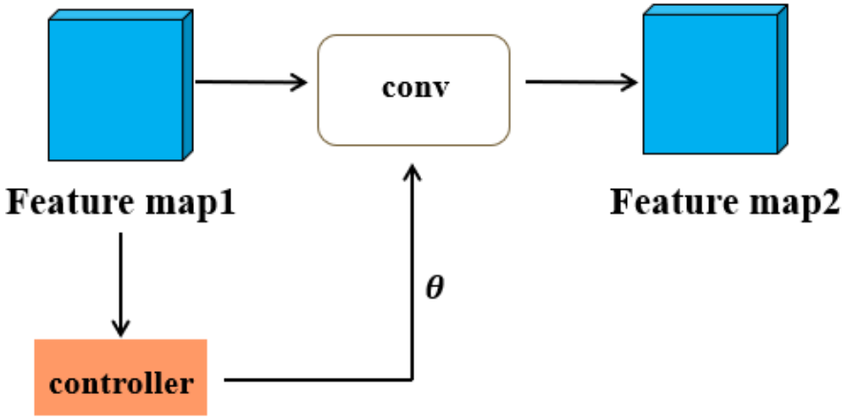
——from MobileNet v2

$$pw_1 : \mathbb{R}^{C_{in} \times \frac{H}{2^i} \times \frac{W}{2^i}} \rightarrow \mathbb{R}^{C_{hidden} \times \frac{H}{2^i} \times \frac{W}{2^i}},$$

$$dw : \mathbb{R}^{C_{hidden} \times \frac{H}{2^i} \times \frac{W}{2^i}} \rightarrow \mathbb{R}^{C_{hidden} \times \frac{H}{2^i} \times \frac{W}{2^i}},$$

$$pw_2 : \mathbb{R}^{C_{hidden} \times \frac{H}{2^i} \times \frac{W}{2^i}} \rightarrow \mathbb{R}^{C_{out} \times \frac{H}{2^i} \times \frac{W}{2^i}}$$





Channel = out_channel × in_channel × k_h × k_w

01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Weight mapper**
  - Employing the full signal in each m_i is inefficient, because φ (Signal feature)is directly mapped into a large number of weights.

m_5 channels	$82 \rightarrow 64$
m_4 channels	$94 \rightarrow 32$
m_3 channels	$44 \rightarrow 16$
m_2 channels	$24 \rightarrow 48 \rightarrow 16$
m_1 channels	$22 \rightarrow 44 \rightarrow 12$

```
Signal feature channel: 1280

Weight_num for each stage: [7298, 4070, 760, 270, 156]
```

01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Weight mapper**
  - divide the channels of φ into parts, $C_{\varphi 0}$, . . . , $C_{\varphi n}$, which are relative in size to the required number of weights of each meta.
  - the number of groups $\mathbf{g_{wi}}$, controls the amount of computations and trainable parameters invested in producing the weights for mi.

Algorithm 1 Divides the channels, C, in unit size, s_u, into chunks relative to the weights, $w_0, \ldots, w_n$.

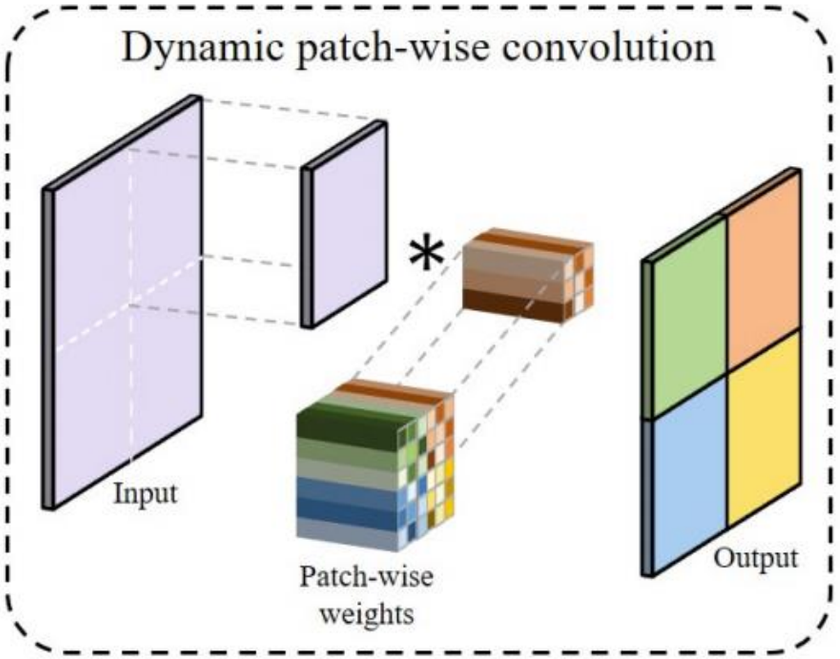
procedure DIVIDE_CHANNELS($C, s_u, w_0, \ldots, w_n$)
  $total\_units \leftarrow \frac{C}{s_u}$
  $w \leftarrow sort(w_0, \ldots, w_n)$ ▷ Descending order
  $r \leftarrow \frac{total\_units}{\sum_{i=0}^{n} w_i}$ ▷ Units to weights ratio
  $out \leftarrow \{s_u | \text{ for each } w_i \in w\}$ ▷ Each weight group should be allocated with at least one unit
  $total\_units \leftarrow total\_units - |out|$
  $i \leftarrow 0$
  while $total\_units \neq 0$ **do**
    if $i = n$
      $curr\_units \leftarrow total\_units$
    else
      $curr\_units \leftarrow \lfloor w_i \cdot r \rfloor - 1$
    $out_i \leftarrow out_i + curr\_units \cdot s_u$
    $total\_units \leftarrow total\_units - curr\_units$
    $i \leftarrow i + 1$
  return out

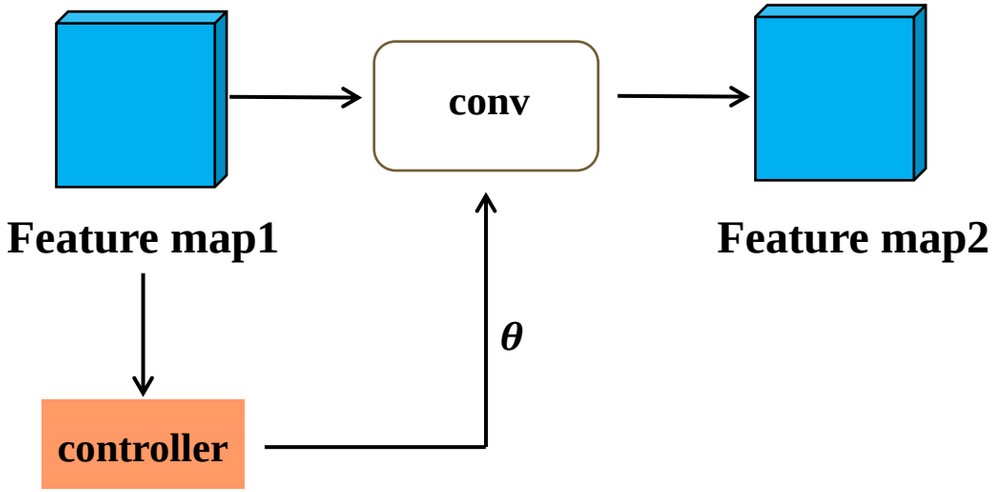
Weight_num for each stage: [7298, 4070, 760, 270, 156]

Feature channel: 1280 ⟶ [736 400 64 16 64]

01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Dynamic patch-wise convolution**
  - Each color represents weights corresponding to a specific patch and '*' is the convolution operation.





before

Batchsize×**Channel** × 1 × 1

Channel = out_channel × in_channel × k_h × k_w

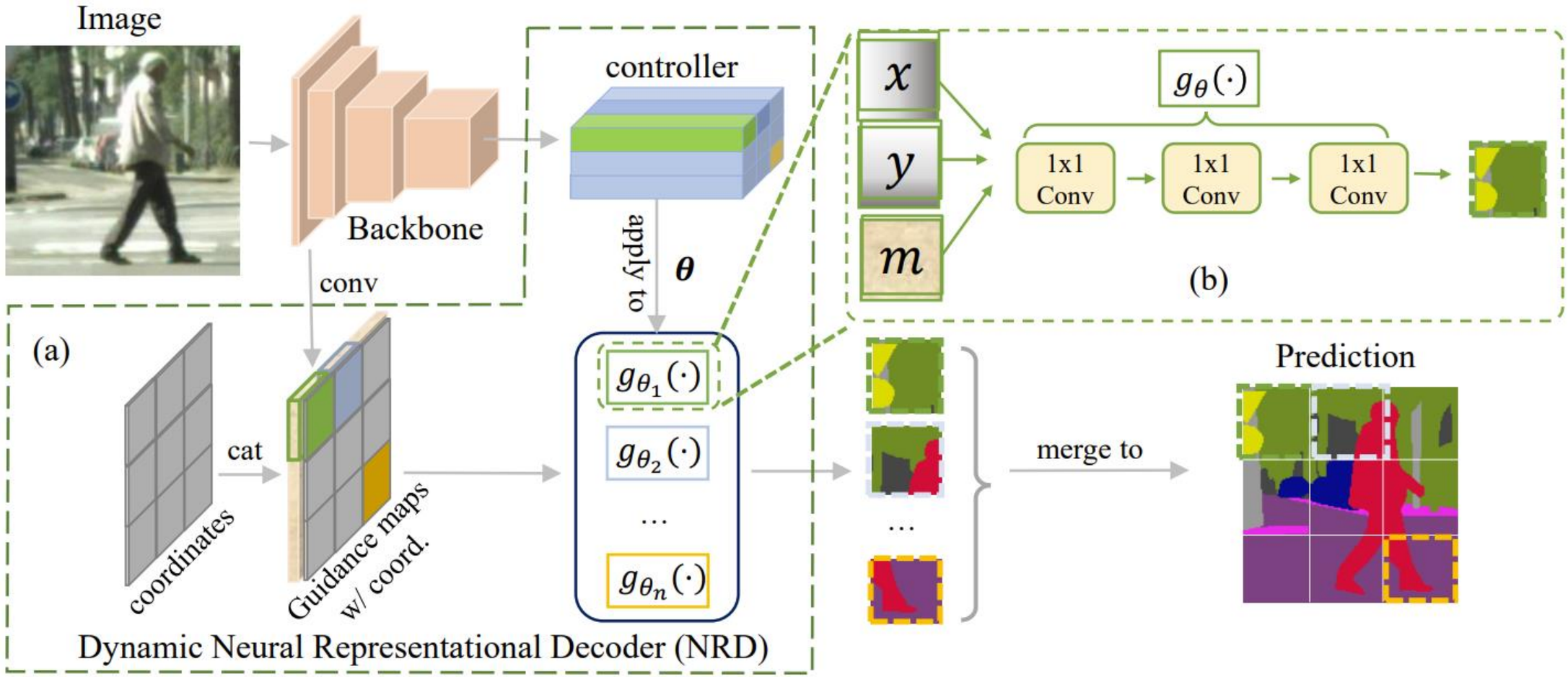
after

Batchsize×**Channel** × ph x pw

Channel = out_channel × in_channel × k_h × k_w

01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Dynamic patch-wise convolution**



Dynamic Neural Representational Decoders for High-Resolution Semantic Segmentation

01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Result-resolution**

Method	Backbone	Resolution	mIoU (%) val	mIoU (%) test	FPS	GFLOPs	Params (M)
ERFNet [38]	-	1024 × 512	-	69.7	41.7	21.7⋆	2.0⋆
ESPNet [30]	ESPNet	1024 × 512	-	60.3	112.9	-	-
ESPNetV2 [30]	ESPNetV2	1024 × 512	66.4	66.2	61.9⋆	2.7	**1.3**⋆
ICNet [57]	PSPNet50	2048 × 1024	-	69.5	30.3	-	-
GUNet [29]	DRN-D-22	1024 × 512	69.6	70.4	33.3	-	-
DFANet A' [25]	Xception A	1024 × 512	-	70.3	**160.0**	**1.7**	7.8
DFANet A [25]	Xception A	1024 × 1024	-	71.3	100.0	3.4	7.8
SwiftNetRN-18 [32]	ResNet18	2048 × 1024	75.4	75.5	39.9	104.0	11.8
BiSeNetV1 [53]	ResNet18	1536 × 768	74.8	74.7	65.5	75.2⋆	49.0
BiSeNetV2 [52]	-	1024 × 512	73.4	72.6	156.0	21.2	-
BiSeNetV2-L [52]	-	1024 × 512	75.8[1]	75.3	47.3	118.5	-
TD4-Bise18 [22]	BiseNet18	2048 × 1024	75.0	74.9	47.6	-	-
HyperSeg-M	EfficientNet-B1	1024 × 512	76.2	75.8	36.9	7.5	10.1
HyperSeg-S	EfficientNet-B1	1536 × 768	**78.2**	**78.1**	16.1	17.0	10.2

Params	HyperSeg-L (PASCAL VOC)	HyperSeg-S (Cityscapes)	HyperSeg-M (Cityscapes)
Backbone	EfficientNet-B3	EfficientNet-B1	EfficientNet-B1
Resolution	512×512	1536×768	1024×512
$r_1, \ldots, r_5$	$1/4, 1/4, 1/4, 1/4, 1/4$	-, $2/5, 1/4, 1/5, 1/6$	$1/4, 1/4, 1/4, 1/4, 1/4$
$g_{w_0}, \ldots, g_{w_5}$	16, 16, 16, 16, 16, 16	-, 4, 16, 8, 16, 32	-, 4, 16, 8, 16, 32
m_5 channels	$98 \rightarrow 96$	$130 \rightarrow 32$	$82 \rightarrow 64$
m_4 channels	$132 \rightarrow 34$	$62 \rightarrow 16$	$94 \rightarrow 32$
m_3 channels	$48 \rightarrow 96 \rightarrow 12$	$26 \rightarrow 8$	$44 \rightarrow 16$
m_2 channels	$22 \rightarrow 44 \rightarrow 8$	$14 \rightarrow 28 \rightarrow 8$	$24 \rightarrow 48 \rightarrow 16$
m_1 channels	$16 \rightarrow 32 \rightarrow 6$	$26 \rightarrow 52 \rightarrow 19$	$34 \rightarrow 68 \rightarrow 19$
m_0 channels	$11 \rightarrow 22 \rightarrow 21$	-	-

01 HyperSeg: Patch-wise Hypernetwork for Real-time Semantic Segmentation

- **Result-backbone/ gride size**

Method	Backbone	mIoU (%)	FPS	Params (M)
ICNet [57]	PSPNet50	67.1	27.8	-
DFANet A [25]	Xception A	64.7	120.0	**7.8**
SwiftNetRN-18 [32]	ResNet18	72.6	85.8*	11.8
BiSeNetV1 [53]	ResNet18	68.7	116.2	49.0
BiSeNetV2 [52]	-	72.4	**124.5**	-
BiSeNetV2-L [52]	-	73.2	32.7	-
TD4-PSP18 [22]	PSPNet18	72.6	25.0	-
TD2-PSP50 [22]	PSPNet50	76.0	11.1	-
HyperSeg-S	ResNet18	77.0	32.5	16.2
HyperSeg-L	ResNet18	77.1	11.5	16.7
HyperSeg-S	PSPNet18	76.6	31.3	17.2
HyperSeg-L	PSPNet18	77.5	11.4	17.6
HyperSeg-S	PSPNet50	77.1	9.3	57.6
HyperSeg-L	PSPNet50	77.9	2.2	67.9
HyperSeg-S w/o DPWConv	EfficientNet-B1	77.3	45.5	9.9
HyperSeg-L w/o DPWConv	EfficientNet-B1	78.4	21.6	10.3
HyperSeg-S	EfficientNet-B1	78.4	38.0	9.9
HyperSeg-L	EfficientNet-B1	**79.1**	16.6	10.2

Grid size	Positional encoding	mIoU (%)	FPS
1×1	✗	77.56	**46.8**
4×4	✗	78.92	22.4
8×8	✗	80.23	26.9
16×16	✗	80.33	28.2
16×16	✓	**80.61**	26.8