A. Detailed experimental settings

A.1. Image classification on ImageNet-1K

We pre-train HRViT on ImageNet-1K for image classification. To generate logits for classification, we append a classification head from HRNetV2 at the end of the HRViT backbone. Four multi-scale outputs from HRViT are fed into convolutional bottleneck blocks, and the channels are mapped to 128, 256, 512, and 1024. Then, we use stride-2 CONV3x3 to down-sample the highest resolution by 2× and double the channels by 2×, and we add it to the next smaller resolution. We repeat this process until we have the smallest resolution features. Finally, we project the 1024-channel feature to 2048 channels via CONV1x1, followed by a global average pooling and a linear classifier.

We adopt a default image resolution of 224×224 and train HRViT with AdamW [22] optimizer for 300 epochs using a cosine learning rate decay schedule, 20 epochs of linear warm-up, and an initial learning rate of 1e-3, a mini-batch size of 1,024, a weight decay rate of 0.05. We set the weight decay rate for BatchNorm layers to 0. We employ gradient clipping with a maximum magnitude of 1. We employ label smoothing with a rate of 0.1 and various data augmentation methods used in DeiT [28], including RandAugment [9], Cutmix [41], Mixup [42], and random erasing [44]. Note that model EMA and repeated augmentation are not employed here. We use a stochastic drop path rate of 0.1 for all backbones. For the classification head in HRViT, we use a dropout rate of 0.1 for all variants. Models are trained on 32 NVIDIA V100 GPUs with 32 images per GPU.

A.2. Semantic segmentation on ADE20K and Cityscapes

ADE20K [45] is a semantic segmentation dataset with 150 semantic categories. It contains 25K images in total, including 20K images for training, 2K for validation, and the rest 3K for testing. Cityscapes [8] dataset contains 5000 fine-annotated high-resolution images with 19 categories. The training image size for ADE20K and Cityscapes are cropped to 512×512 and 1024×1024, respectively. For data augmentation, we use random horizontal flipping, random re-scaling within the ratio range of [0.5, 2.0], and random photometric distortion. On ADE20K, the stochastic drop path rates are set to 0.1, 0.1, 0.25 for HRViT-b1, HRViT-b2, and HRViT-b3, respectively. On Cityscapes, the stochastic drop path rates are set to 0.15, 0.15, 0.25 for HRViT-b1, HRViT-b2, and HRViT-b3, respectively. We use an AdamW optimizer for 160 k iterations using a 'poly' learning rate schedule, 1,500 steps of linear warm-up, an initial learning rate of 6e-5, and a weight decay rate of 0.01. The mini-batch size is set to 16 and 8 for ADE20K and Cityscapes, respectively. We set the weight decay rate for BatchNorm layers to 0 and increase the initial learning rate for the segmentation head by 10×. Models on ADE20K are trained on 8 NVIDIA V100 GPUs with 2 images per GPU. Models on Cityscapes are trained on 8 NVIDIA V100 GPUs with 1 image per GPU. During inference, the image size for ADE20K val and Cityscapes val is set to 512×2048 and 1024×2048, respectively. We do inference on Cityscapes with sliding window test by cropping 1024×1024 patches.

B. More experiments

B.1. Different block assignment strategies

Block Assignment on the 3rd Branch	ImageNet-1K Top-1 Acc	Cityscapes val mIoU
6-6-6-2	80.53	81.63
8-8-2-2	80.51	81.50
9-9-1-1	80.50	81.25
17-1-1-1	80.11	81.27

Table 8. Compare different block assignment strategies on the third LR branch in HRViT-b1.

In Table 8, we compare different block assignment strategies on the 3rd low-resolution path in HRViT-b1 on ImageNet-1K and Cityscapes. We observe a clear trend that when concentrating more blocks in one module, e.g., from 6-6-6-2 to 17-1-1-1, the benefits from the cross-resolution fusion in the HR architecture diminish accordingly, leading to degraded ImageNet classification accuracy and Cityscapes segmentation mIoU. This phenomenon can be attributed to the information loss in the deep LR module, which validates the effectiveness of our *even block assignment* strategy with better information interaction and detail preservation.

B.2. Semantic segmentation on ADE20K with UperNet head

Besides the lightweight SegFormer head, UperNet [33] is another segmentation framework that is widely used. We evaluate HRViT on ADE20K val with SegFormer and UperNet head in Table 9. Our HRViT can mostly maintain the advantages, but the computation benefits are not as significant as with the SegFormer head. The reason is that the UperNet head dominates the computations (>89%) and parameters (>54%) in the cost breakdown. Hence any slimming on backbones will be considerably diluted. Moreover, we observe similar performance with lightweight SegFormer head and heavy UperNet head on HRViT. Moreover, we do not observe more performance benefits from UperNet head than the SegFormer head on HRViT. One explanation is that *HRViT already has enough multi-resolution fusion, which makes the additional pyramid fusion in UperNet head less effective than used in the*

Backbone	SegFormer Head [35]			UperNet Head [33]		
	#Param. (M)↓	GFLOPs↓	mIoU (%)↑	#Param. (M)↓	GFLOPs↓	mIoU (%)↑
MiT-B0 [35]	3.8	8.4	37.40	-	-	-
MiT-B1 [35]	13.7	15.9	42.20	-	-	-
CSWin-Ti [12]	5.9	11.4	41.43	33.7	216	44.41
HRViT-b1	**8.2**	**14.6**	**45.88**	**35.9**	**219**	**47.19**
Twins-S [6]	-	-	-	54.4	219	46.20
Swin-T [20]	-	-	-	59.9	234	44.50
MiT-B2 [35]	27.5	62.4	46.50	-	-	-
CSWin-T [12]	22.4	28.3	47.88	59.9	234	49.30
HRViT-b2	**20.8**	**28.0**	**48.76**	**49.7**	**233**	**49.10**
Twins-B [6]	-	-	-	88.5	248	47.70
Swin-S [20]	-	-	-	81.3	253	47.60
MiT-B3 [35]	47.3	79.0	49.40	-	-	-
CSWin-S [12]	37.3	78.1	49.93	64.6	247	50.00
HRViT-b3	**28.7**	**67.9**	**50.20**	**55.4**	**236**	**50.04**
Avg improv.	-25.2%	-19.7%	+2.06	-25.6%	-2.7%	+1.54

Table 9. Performance and efficiency comparison of different ViT backbones on the ADE20K val segmentation dataset. Average improvements of HRViT over baselines are summarized for each framework.

sequential architectures. Hence, the lightweight SegFormer head is more suitable to HRViT.