DVHGNN: Multi-Scale Dilated Vision HGNN for Efficient Vision Recognition

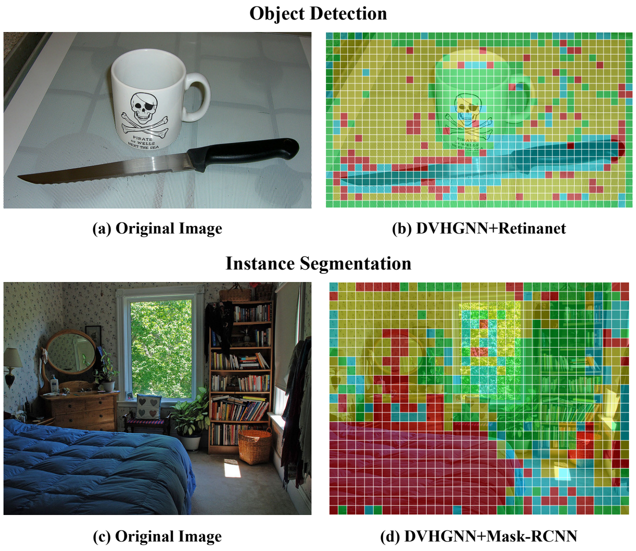
Supplementary Material

A. Cross-task Visualizations

In Figure 1, we present a visualization of the clustering hyperedges extracted from a single attention head in the final stage of DVHGNN-S, where the feature map size is set to 25×38. Specifically, Figure 1 (b) and Figure 1 (d) illustrate the clustering results for object detection and instance segmentation tasks, respectively, under the RetinaNet and Mask R-CNN architectures. To facilitate interpretation, different hyperedge types are distinguished by color, encoding the underlying semantic relationships among visual entities. For example, in Figure 1 (b), the blue hyperedges correspond to the knife category, while in Figure 1 (d), the red hyperedges represent the bed category. Notably, we omit the visualization of dilated hyperedges, as their structural configurations remain consistent with those elaborated in the main paper. This visualization further substantiates the efficacy of our hypergraph-based approach in capturing meaningful semantic structures across diverse tasks.

B. Additional Ablation Study

Impact of the Clustering Step on Inference Speed. Since the clustering method constitutes the fundamental backbone of our approach, direct ablation analysis on this component is not feasible. Therefore, we instead assess the impact of the proposed Dynamic Hypergraph Convolution (DHGC) and Dynamic Hypergraph Convolutional Filtering (DHConv) on inference efficiency. As reported in Table 1, removing DHGC leads to an inference speed improvement of 0.084 ms per image, while removing DHConv further accelerates inference by 0.087 ms per image. However, these speed gains come at the cost of a noticeable degradation in Top-1 accuracy, with DHGC and DHConv ablations resulting in 0.8% and 1.1% drops in performance, respectively. These results underscore the inherent trade-off between computational efficiency and predictive accuracy. While eliminating these components marginally improves throughput, the corresponding accuracy degradation suggests that DHGC and DHConv play a crucial role in enhancing model representation capacity and overall performance. Therefore, the slight increase in computational overhead is justified by the substantial accuracy gains, reinforcing the effectiveness of our proposed hypergraph-based feature aggregation strategy.



Object Detection

(a) Original Image

(b) DVHGNN+Retinanet

Instance Segmentation

(c) Original Image

(d) DVHGNN+Mask-RCNN

Figure 1. Visualization of cross-tasks clustering hyperedges in object detection and instance segmentation, respectively.

Method	Params	FLOPs	Throughput	Top-1
DVHGNN-T$_{\text{w/o ConvFFN}}$	11.1 M	1.8 G	874.5 img/ms	78.0%
w/o DHGC	11.1 M	1.8 G	943.8 img/ms	77.2%
w/o DHConv	11.0 M	1.7 G	1028.1 img/ms	76.1%

Table 1. Impact of distinct moudule on inference throughput.

Models	Channels	Blocks	Heads	D'
DVHGNN-T	[48, 96, 240, 480]	[2, 2, 6, 2]	[3, 6, 12, 24]	24
DVHGNN-S	[64, 128, 320, 640]	[3, 3, 9, 3]	[4, 8, 16, 32]	32
DVHGNN-M	[96, 192, 384, 768]	[4, 4, 14, 4]	[4, 8, 16, 32]	32
DVHGNN-T	[96, 192, 384, 768]	[6, 6, 24, 6]	[5, 10, 20, 40]	32

Table 2. Configurations of different DVHGNN variants.