TaskExpert: Dynamically Assembling Multi-Task Representations with Memorial Mixture-of-Experts

-Supplemental Material-

Hanrong Ye and Dan Xu
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology (HKUST)
hyeae@cse.ust.hk, danxu@cse.ust.hk

1. Additional Implementation Details

This section provides additional implementation details regarding the data processing and the model optimization for our proposed model.

Data Processing We follow the data processing pipeline of InvPT [5] for a fair comparison. Specifically, on NYUD-v2 dataset, we crop the input image to the size of 448×576 randomly. On PASCAL-Context dataset, we pad the image to a size of 512×512. We also use the same data augmentation techniques, including random scaling, color jittering, cropping, and horizontal flipping.

Model Optimization We totally investigate 6 different tasks on the challenging NYUD-v2 [3] and PASCAL-Context [1] benchmarks, including semantic segmentation (Semseg), monocular depth estimation (Depth), surface normal estimation (Normal), human parsing (Parsing), saliency detection (Saliency), and object boundary detection (Boundary). For the continuous regression tasks, such as Depth and Normal, we employ the $\mathcal{L}1$ Loss. For the discrete classification tasks, including Semseg, Parsing, Saliency, and Boundary, we utilize the cross-entropy loss. The overall loss is a weighted sum of the task losses from all the tasks, using a loss weight dictionary utilized by [4,5]. More specifically, the loss weight is 1 for Semseg, 2 for Parsing, 5 for Saliency, 50 for Boundary, 1 for Depth, and 10 for Normal.

2. More Study of Gating Networks

2.1. Ablation Study of Kernel Sizes in Context-Aware Gating Networks

Context-Aware Gating is a critical module in MMoE that incorporates contextual information into the computation of the gating score for each token. In this experimental study, we aim to study the influence of kernel sizes used in convolutions of the gating networks in Table 1. Compared with the variant using 1×1 convolution kernel size which resembles an MLP, using 3×3 clearly brings improvement on the performances of all the tasks. For instance, the mIoU of Semseg and Parsing is improved from 77.73 and 66.43 to 78.45 (+0.72) and 67.38 (+0.95), respectively. However, further increasing the kernel size does not lead to further improvement on all the tasks. This is because the most important information for computing the gating score of a given token typically comes from spatially adjacent tokens. For this reason, we have chosen to use a default convolution kernel size of 3×3 in our context-aware gating module.

Table 1: Quantitative comparison of using different sizes of convolution kernels in the gating networks. 3×3 is large enough to perceive the context of each token for gating networks. It clearly outperforms the variant with 1×1 kernel on all the tasks, and achieves the best performance on most of the tasks compared to all the variants.

Kernel Size	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑
1×1	77.73	66.43	84.67	13.78	71.00
3×3	**78.45**	**67.38**	84.96	13.55	**72.30**
5×5	78.40	66.59	**85.22**	**13.48**	72.10

2.2. Comparison with MLP Gating Network

A standard Mixture-of-Experts (MoE) gating network typically incorporates a simplistic linear layer. To establish a balanced comparison, ensuring a comparable parameter size, we implement a conventional Multi-Layer Perceptron (MLP) consisting of two linear layers as the gating network, without incorporating the multi-task feature memory mechanism. The corresponding results, presented in Table 2, highlight that our context-aware gating (CG) strategy achieves superior performance, while simultaneously consuming fewer parameters.

Table 2: Ablation study of gating networks on PASCAL-Context. "CG" denotes the proposed context-aware gating.

Model	Semseg mIoU ↑	Parsing mIoU ↑	Saliency maxF ↑	Normal mErr ↓	Boundary odsF ↑	MTL Perf Δ_m ↑	#Param
MLP-Gating	77.13	66.25	84.64	13.80	70.80	-2.96	256M
CG	77.99	66.82	84.76	13.64	71.70	-2.07	227M

3. More Comparison with Previous SOTA

To confirm the superior performance of the proposed TaskExpert, we visualize its task-specific prediction maps on the testing set, and compare them with the output of the previous best state-of-the-art method (*i.e*. InvPT [5]) on the PASCAL-Context [2] and NYUD-v2 [3] datasets. The results are presented in Figure 1 and Figure 2, respectively. The qualitative study demonstrates that our TaskExpert can generate finer results, particularly in semantic segmentation and human parsing.

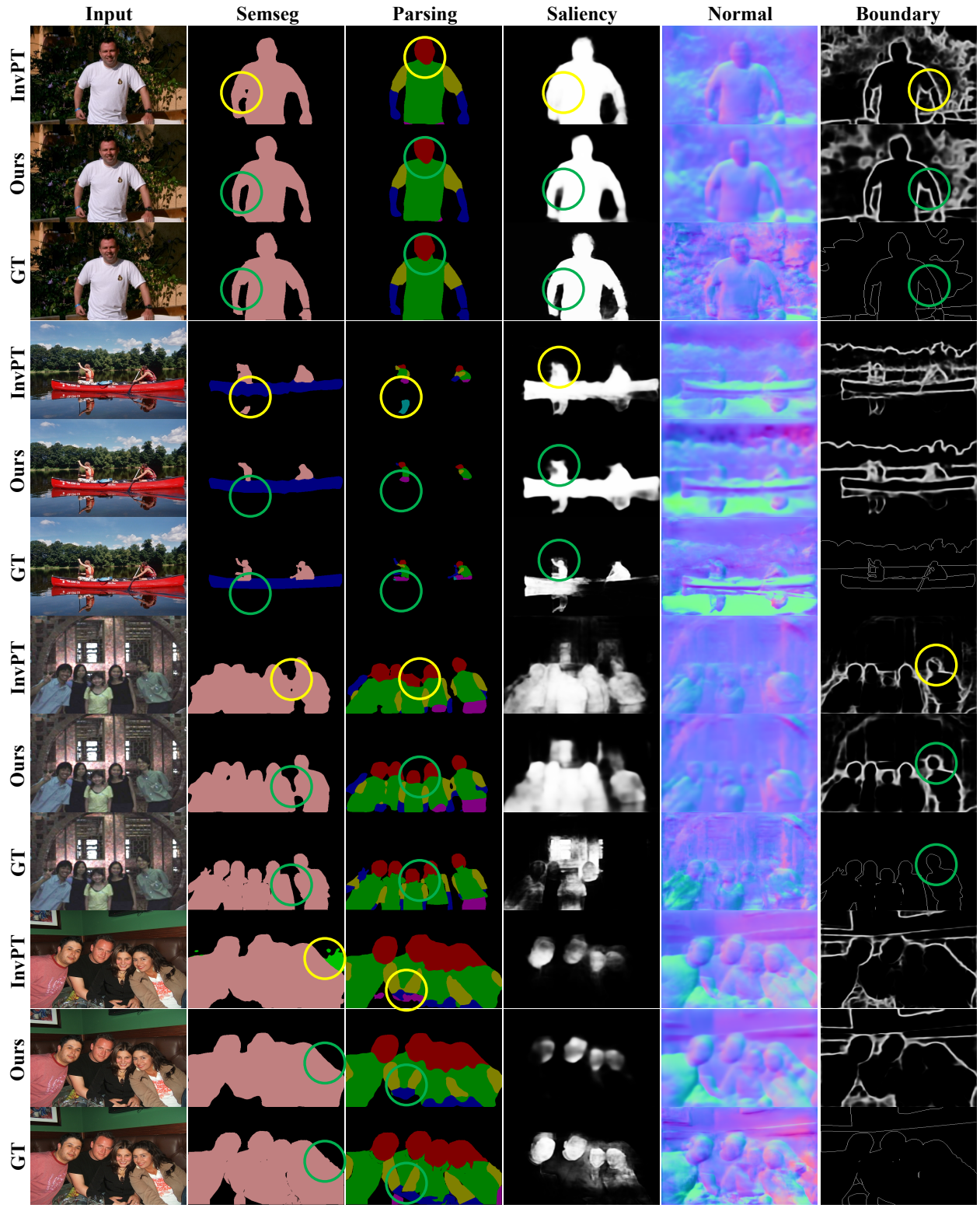

Figure 1: Qualitative comparison with previous best-performing method InvPT [5] on PASCAL-Context. Our method generates significantly better results, especially on semantic segmentation and human parsing, as highlighted in circles.

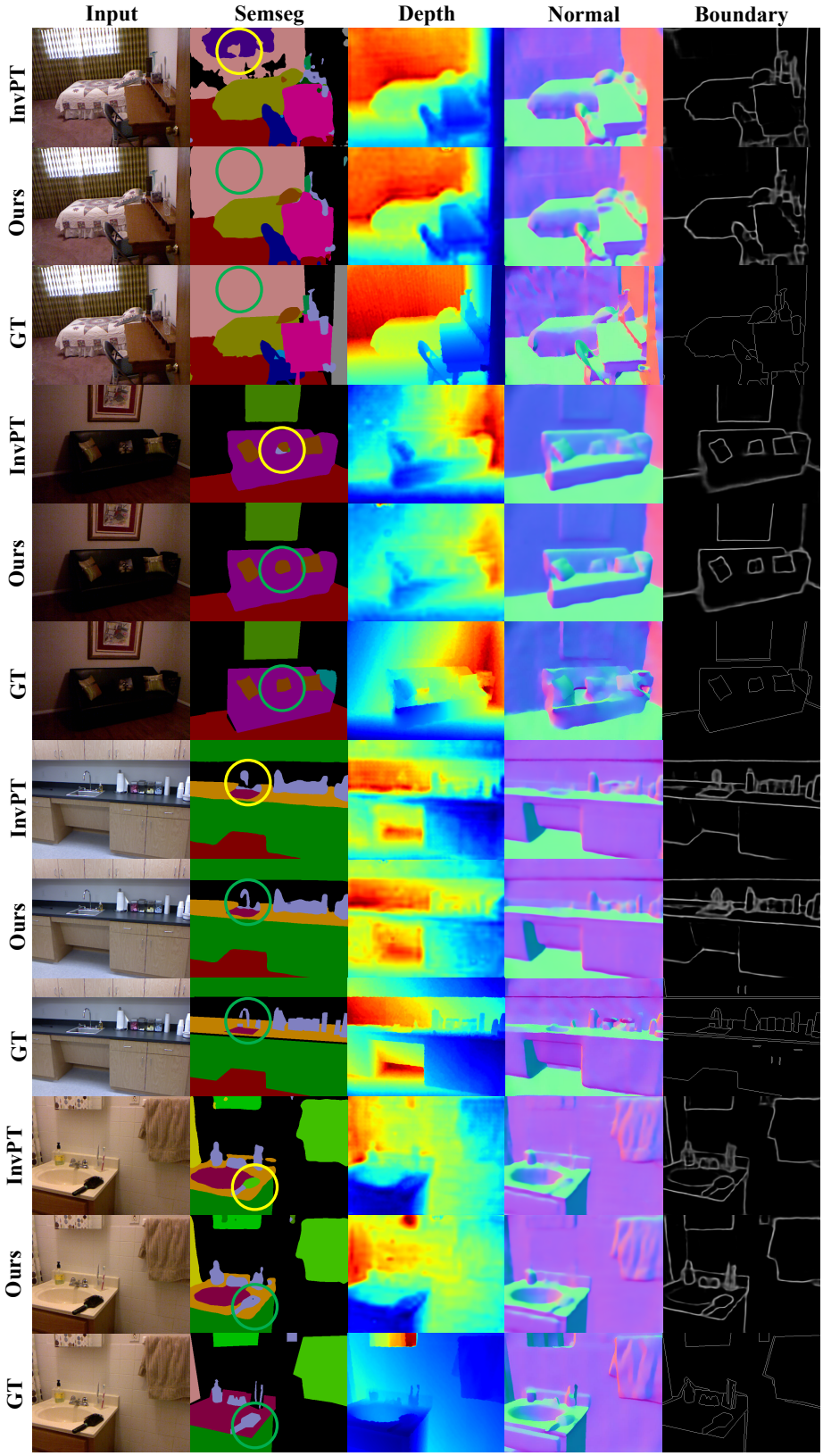

Figure 2: Qualitative comparison with the previous best-performing method InvPT [5] on NYUD-v2. Our method generates significantly better results than previous best-performing InvPT, as highlighted in circles.

References

[1] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 1

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 2

[3] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 2

[4] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020. 1

[5] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022. 1, 2, 3, 4