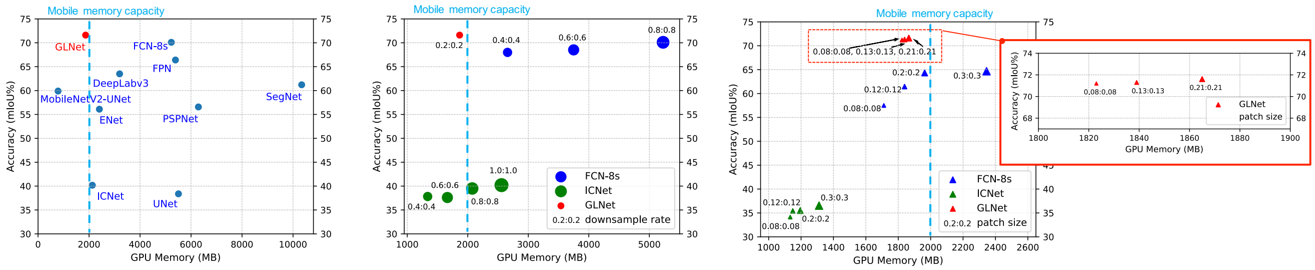
# Supplementary Materials: Collaborative Global-Local Networks for Memory-Efficient Segmentation of Ultra-High Resolution Images

Wuyang Chen*[1], Ziyu Jiang*[1], Zhangyang Wang[1], Kexin Cui[1] and Xiaoning Qian[2]

{*wuyang.chen, jiangziyu, atlaswang, ckx9411sx*}*@tamu.edu, xqian@ece.tamu.edu*
[1]Department of Computer Science and Engineering, Texas A&M University
[2]Department of Electrical and Computer Engineering, Texas A&M University
https://github.com/chenwydj/ultra_high_resolution_segmentation

## 1. Additional Results

In this supplementary material we added the ablation study of the **ENet** [1] and **MobileNetV2-UNet** [2, 3] on the DeepGlobe dataset [4], since they are very efficient off-the-shelf backbones and their designs considered the accuracy-efficiency trade-off. Table 1 list a complete ablation study of mIoU and memory usage comparison on the DeepGlobe dataset. From Table 1 we can see that all models achieve higher mIoU under global inference, but consume very high GPU memories. Their memory usages will drop if adopting patch-based inference, but accuracies also deteriorate accordingly. Our GLNet achieves the best trade-off between mIoU and GPU memory usage. We also included a **detailed comparison of the performance of our GLNet with different patch sizes** in Figure 1 (c), where the zoom-in panel shows that the accuracy of our GLNet is highly preserved under different patch sizes, and the GPU memory usage has the minimum changes comparing to the FCN-8s[5] and the ICNet [6].



(a) Best performance achievable

(b) Performance trained on global image

(c) Performance trained on local patches

**Figure 1:** Inference memory and mean intersection over union (mIoU) accuracy on the DeepGlobe dataset [4]. (a): Comparison of best achievable mIoU v.s. memory for different segmentation methods. (b): mIoU/memory with different global image sizes (downsampling rate shown in annotations). (c): mIoU/memory with different local patch sizes (normalized patch size shown in annotations), with the zoom-in panel showing the performance of the GLNet with different patch sizes. **GLNet** (red dots) integrates both global and local information in a compact way, contributing to a well-balanced trade-off between accuracy and memory usage. See Section 4.3 for experiment details. Methods studied: ICNet [6], DeepLabv3+ [7], FPN [8], FCN-8s [5], UNet [9], PSPNet [10], SegNet [11], ENet [1], MobileNetV2-UNet [2, 3], and the proposed GLNet.

We have chosen several state-of-the-art models with public implementations for comparison on the DeepGlobe [4], ISIC [12, 13], and Inria Aerial [14] Datasets. These datasets have segmentation performance leaderboards, but most leading models either do not disclose full technical details, or rely on heavily parameterized ensemble models for accuracy (thus unfair to compare with). Furthermore, few models on leaderboards open-sourced their implementations, making us unable to test their

*The first two authors contributed equally.

**Table 1:** Predicted mIoU and inference memory usage on the local DeepGlobe test set. '$\mathcal{G} \rightarrow \mathcal{L}$' and '$\mathcal{G} \rightleftarrows \mathcal{L}$' means feature map sharing from the global to local branch and bidirectionally between two branches respectively. Note that our GLNet does not inference with global images.

| Model | Patch Inference | | Global Inference | |
|---|---|---|---|---|
| | mIoU(%) | Memory(MB) | mIoU(%) | Memory(MB) |
| UNet[9] | 37.3 | 949 | 38.4 | 5507 |
| ICNet[6] | 35.5 | 1195 | 40.2 | 2557 |
| PSPNet[10] | 53.3 | 1513 | 56.6 | 6289 |
| SegNet[11] | 60.8 | 1139 | 61.2 | 10339 |
| DeepLabv3+[7] | 63.1 | 1279 | 63.5 | 3199 |
| FCN-8s[5] | 64.3 | 1963 | 70.1 | 5227 |
| ENet[1] | 55.5 | 680 | 56.1 | 2405 |
| MobileNetV2-UNet[2, 3] | 59.9 | 785 | 54.5 | 1797 |
| | mIoU(%) | | Memory(MB) | |
| **GLNet:** $\mathcal{G} \rightarrow \mathcal{L}$ | **70.9** | | **1395** | |
| **GLNet:** $\mathcal{G} \rightleftarrows \mathcal{L}$ | **71.6** | | **1865** | |

GPU memory usages. Instead, we notice that most single models on leaderboards are modified from latest backbones like UNet, FCN, and FPN, with standard implementations available. Therefore, we fine-tune those backbones's performance on each dataset, and make them as our comparision subjects. Also because the challenge testing sets are not publicly available, we make fair comparisons on our own training-testing split, unless otherwise stated.

## 2. Training Strategy in Details

We depict our training strategy details in the Algorithm 1. '$\mathcal{G} \rightarrow \mathcal{L}$' stands for the deep feature map from the global to local branch, and '$\mathcal{G} \rightleftarrows \mathcal{L}$' means the deep feature map sharing bidirectionally between two branches. The '$Focal$' stands for the Focal Loss [15] we used in our experiments with $\gamma = 6$.

**Algorithm 1** Collaborative Global-Local Networks

```
Input: Ultra-high resolution images and segmentation maps $\mathcal{D} = \{(\boldsymbol{I}_i, \boldsymbol{S}_i)\}_{i=1}^{N}$ where $\boldsymbol{I}_i, \boldsymbol{S}_i \in \mathbb{R}^{H \times W}$, down-sampled low resolution images and segmentation maps $\mathcal{D}^{\text{lr}} = \{(\boldsymbol{I}_i^{\text{lr}}, \boldsymbol{S}_i^{\text{lr}})\}_{i=1}^{N}$ where $\boldsymbol{I}_i^{\text{lr}}, \boldsymbol{S}_i^{\text{lr}} \in \mathbb{R}^{h_1 \times w_1}$, cropped image and segmentation map patches $\mathcal{D}^{\text{hr}} = \{\{(\boldsymbol{I}_{ij}^{\text{hr}}, \boldsymbol{S}_{ij}^{\text{hr}})\}_{j=1}^{n_i}\}_{i=1}^{N}$, where each $\boldsymbol{I}_i$ and $\boldsymbol{S}_i$ in $\mathcal{D}$ comprises $n_i$ patches and $\boldsymbol{I}_i^{\text{hr}}, \boldsymbol{S}_i^{\text{hr}} \in \mathbb{R}^{h_2 \times w_2}$. $h_1, h_2 \ll H$, and $w_1, w_2 \ll W$
Output:
Initialization: global branch $\mathcal{G} = f_{\text{clf}}^{\mathcal{G}} \circ f_{\text{feature}}^{\mathcal{G}}$, local branch $\mathcal{L} = f_{\text{clf}}^{\mathcal{L}} \circ f_{\text{feature}}^{\mathcal{L}}$, where $f_{\text{feature}}^{\mathcal{G}}, f_{\text{feature}}^{\mathcal{L}}$ each has L layers; Aggregation layer $f_{\text{agg}}$
Train $\mathcal{G}$ on $\mathcal{D}^{\text{lr}}$:
    $\hat{\boldsymbol{X}}_{i,1}^{\mathcal{G}}, \ldots, \hat{\boldsymbol{X}}_{i,L}^{\mathcal{G}} = f_{\text{feature}}^{\mathcal{G}}(\boldsymbol{I}_i^{\text{lr}})$
    $\hat{\boldsymbol{S}}_i^{\mathcal{G}} = f_{\text{clf}}^{\mathcal{G}}(\hat{\boldsymbol{X}}_{i,L}^{\mathcal{G}})$
    $\min_{f_{\text{feature}}^{\mathcal{G}}, f_{\text{clf}}^{\mathcal{G}}} = \frac{1}{n} \sum_{i=1}^{N} Focal(\hat{\boldsymbol{S}}_i^{\mathcal{G}}, \boldsymbol{S}_i^{\text{lr}})$
$\mathcal{G} \rightarrow \mathcal{L}$: Train $\mathcal{L}$ on $\mathcal{D}^{\text{hr}}$:
    $\hat{\boldsymbol{X}}_{i,1}^{\mathcal{L}}, \ldots, \hat{\boldsymbol{X}}_{i,L}^{\mathcal{L}} = f_{\text{feature}}^{\mathcal{L}}(\boldsymbol{I}_i^{\text{hr}}; \hat{\boldsymbol{X}}_{i,1}^{\mathcal{G}}, \ldots, \hat{\boldsymbol{X}}_{i,L-1}^{\mathcal{G}})$
    $\hat{\boldsymbol{S}}_i^{\mathcal{L}} = f_{\text{clf}}^{\mathcal{L}}(\hat{\boldsymbol{X}}_{i,L}^{\mathcal{L}})$, $\hat{\boldsymbol{S}}_i^{\text{Agg}} = f_{\text{agg}}(\hat{\boldsymbol{X}}_{i,L}^{\mathcal{L}}, \hat{\boldsymbol{X}}_{i,L}^{\mathcal{G}})$
    $\min_{f_{\text{feature}}^{\mathcal{L}}, f_{\text{clf}}^{\mathcal{L}}, f_{\text{agg}}} \frac{1}{n} \sum_{i=1}^{N} Focal(\hat{\boldsymbol{S}}_i^{\mathcal{L}}, \boldsymbol{S}_i^{\text{hr}}) + Focal(\hat{\boldsymbol{S}}_i^{\text{Agg}}, \boldsymbol{S}_i^{\text{hr}}) + \lambda \|\hat{\boldsymbol{X}}_{i,L}^{\mathcal{G}} - \hat{\boldsymbol{X}}_{i,L}^{\mathcal{L}}\|_2$
$\mathcal{G} \rightleftarrows \mathcal{L}$: Train $\mathcal{G}$ on $\mathcal{D}^{\text{lr}}$:
    $\hat{\boldsymbol{X}}_{i1}^{\mathcal{G}}, \ldots, \hat{\boldsymbol{X}}_{iL}^{\mathcal{G}} = f_{\text{feature}}^{\mathcal{G}}(\boldsymbol{I}_i^{\text{lr}}; \hat{\boldsymbol{X}}_{i,1}^{\mathcal{L}}, \ldots, \hat{\boldsymbol{X}}_{i,L-1}^{\mathcal{L}})$
    $\hat{\boldsymbol{S}}_i^{\mathcal{G}} = f_{\text{clf}}^{\mathcal{G}}(\hat{\boldsymbol{X}}_{i,L}^{\mathcal{G}})$, $\hat{\boldsymbol{S}}_i^{\text{Agg}} = f_{\text{agg}}(\hat{\boldsymbol{X}}_{i,L}^{\mathcal{G}}, \hat{\boldsymbol{X}}_{i,L}^{\mathcal{L}})$
    $\min_{f_{\text{feature}}^{\mathcal{G}}, f_{\text{clf}}^{\mathcal{G}}, f_{\text{agg}}} \frac{1}{n} \sum_{i=1}^{N} Focal(\hat{\boldsymbol{S}}_i^{\mathcal{G}}, \boldsymbol{S}_i^{\text{hr}}) + Focal(\hat{\boldsymbol{S}}_i^{\text{Agg}}, \boldsymbol{S}_i^{\text{hr}})$
Return $\hat{\boldsymbol{S}}^{\text{Agg}}$
```

# References

[1] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

[2] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[3] Akira Sosa. Real-time semantic segmentation in mobile device. *https://github.com/akirasosa/mobile-semantic-segmentation*, 2017.

[4] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raska. Deepglobe 2018: A challenge to parse the earth through satellite images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–17209. IEEE, 2018.

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[6] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–420, 2018.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[8] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[10] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[11] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[12] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.

[13] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 168–172. IEEE, 2018.

[14] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017.

[15] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.