nnWNet: Rethinking the Use of Transformers in Biomedical Image Segmentation and Calling for a Unified Evaluation Benchmark

Supplementary Material

1. Motivation of Architecture Design

In practical large-scale applications, we find that the two-stage cascaded UNet can further improve performance. Therefore, we try to expand the U-shaped architecture to W-shape to simulate and achieve the cascade effect. In order to capture long-range dependencies while focusing on local details, we introduce Global Scope Bridges (GSBs) at each scale to achieve a continuous and stable flow of local and global features. Because the W-shaped architecture directly transmits high-resolution features between input and output and maintains multi-scale features of low, medium, and high resolutions, it performs better. By the way, our design concept is similar to HRNet.

2. Datasets

We evaluate our model on four 2D datasets (DRIVE, ISIC-2107, Kvasir-SEG, and CREMI) and four 3D datasets (Parse2022, AMOS22, BTCV, and ImageCAS). Here are their details.

DRIVE. This is a retinal vessel segmentation dataset. It contains 40 fundus images of size 584×565. Each image is annotated by two human observers under the guidance of experienced ophthalmologists, where the first observer segmentation is accepted as the ground truth for performance evaluation.

ISIC-2107. This is a skin lesion segmentation dataset. It contains 2750 dermoscopic images. Each image is annotated by an expert clinician through a semi-automatic or manual process.

Kvasir-SEG. This is a gastrointestinal polyp segmentation dataset. It contains 1000 polyp images with sizes ranging from 332×487 to 1920×1072. All images are manually annotated and verified by an experienced gastroenterologist.

CREMI. This is an electron microscopy dataset for neuronal membrane segmentation. It consists of three image stacks for three different types of neurons. Each stack consists of 125 slices of size 1250×1250.

Parse2022. This is a pulmonary artery segmentation dataset. It contains 100 contrast enhanced CT pulmonary angiography (CTPA) images. The image size ranges from 512×512×228 to 512×512×376. The pixel size of the slices ranges from $0.5mm/pixel$ to $0.95mm/pixel$, and the slice thickness is $1mm/pixel$. Ten experts with more than five years of clinical experience participated in the annotation. The annotation is performed on the basis of region growing algorithm using MIMICS software.

AMOS22. This is an abdominal multi-organ segmentation dataset. It contains 300 abdominal CT images and 60 abdominal MRI images. Five junior radiologists perform semi-automatic annotations on the 15 organs (spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus). Three senior radiologists with more than ten years of clinical experience perform the final calibration.

BTCV. This is an abdominal multi-organ segmentation dataset. It contains 30 abdominal CT images with sizes ranging from 512×512×85 to 512×512×198. The pixel size of the slices ranges from $0.54mm \times 0.54mm$ to $0.98mm \times 0.98mm$, and the slice thickness ranges from $2.5mm$ to $5.0mm$. Thirteen abdominal organs (spleen,

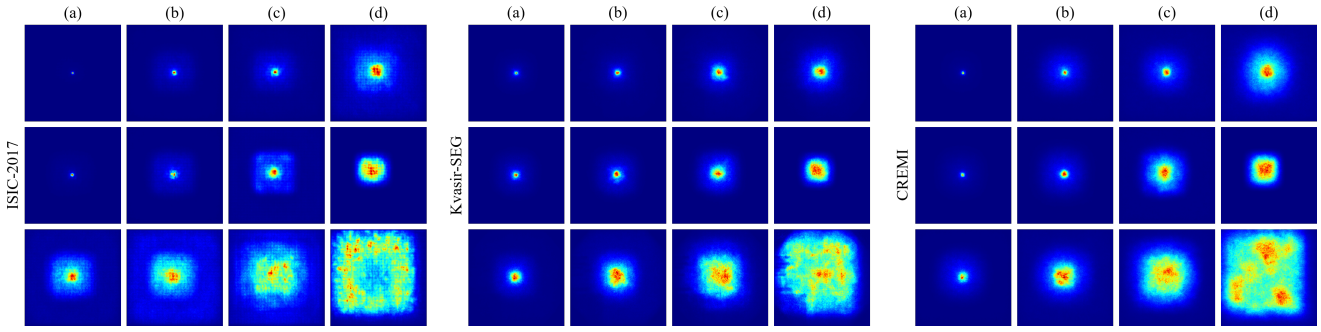

Figure 1. Effective receptive fields (ERFs) of LSBs and GSBs on ISIC2017, Kvasir-SEG, CREMI (average over 100 images). Top row: The ERFs of residual blocks of LSBs in the second encoder. Middle row: The ERFs of residual blocks of LSBs in the first decoder. Botton row: The ERFs of 11×11 depth-wise convolution self-attentions of GSBs between the first decoder and the second encoder. (a) Scale 1. (b) Scale 2. (c) Scale 3. (d) Scale 4.

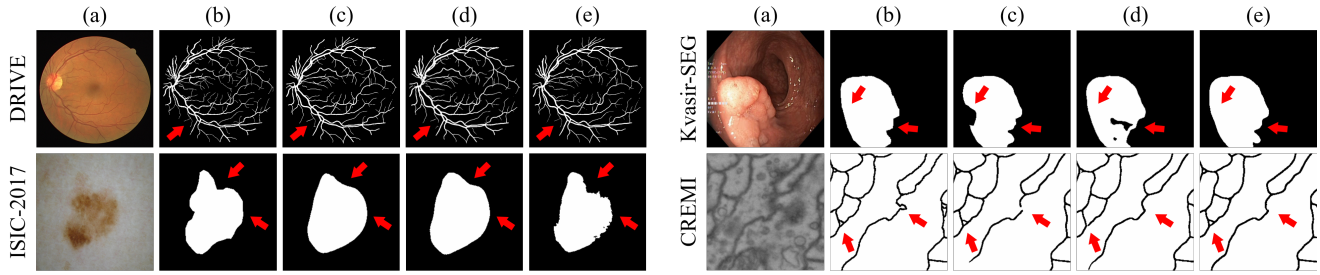

Figure 2. More qualitative results of different models on 2D datasets. (a) Raw images. (b) Ground truth. (c) TransAttUNet. (d) nnUNet. (e) nn~~W~~Net. The red arrows highlight the differences among the results

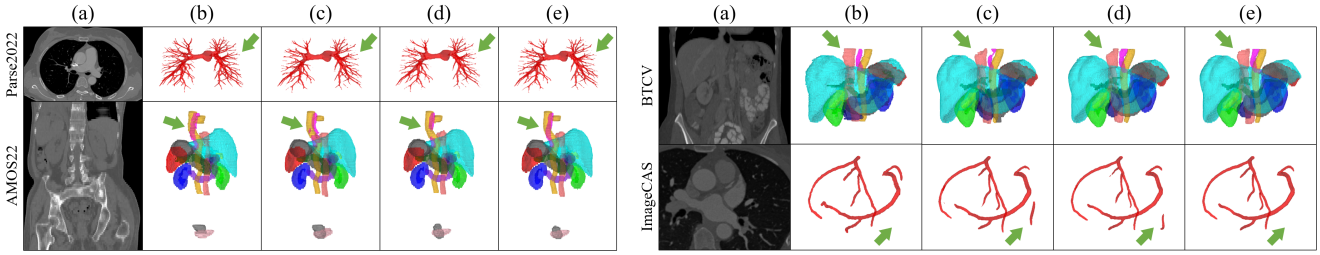

Figure 3. More qualitative results of different models on 3D datasets. (a) Raw images. (b) Ground truth. (c) CoTr. (d) nnUNet. (e) nn~~W~~Net. The green arrows highlight the differences among the results

right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal & splenic vein, pancreas, right adrenal gland, left adrenal gland) are manually annotated by two experienced undergraduate students, and verified by a radiologist on a volumetric basis using the MIPAV software.

ImageCAS. This is a coronary artery segmentation dataset. It contains 1000 computed tomography angiography (CTA) images with sizes ranging from 512×512×206 to 512×512×275. The pixel size of the slices ranges from $0.29mm^2$ to $0.43mm^2$, and the slice thickness ranges from $0.25mm$ to $0.45mm$. Each image is annotated by three radiologists.

To balance training time and effect, we resize the raw images in ISIC-2017 and Kvasir-SEG to 256×256, and use a sliding window to crop the raw images in CREMI to 256×256.

3. Effective Receptive Fields of LSBs and GSBs at Each Scale

We compare the effective receptive fields (ERFs) of LSBs and GSBs in Figure 1. As the scale increases, the ERFs of both LSBs and GSBs expand. At the same scale, the ERF of LSB is significantly smaller than that of GSB. Notably, at scale 4, the ERFs of the two become complementary, with regions in LSB appearing redder, while the corresponding regions in GSB appear bluer. This is because LSBs are composed of convolutions and focus more on local details, whereas GSBs are composed of transformers and tend to capture long-range dependencies.

4. More Qualitative Results

Figure 2 and Figure 3 show more qualitative results. As described in Section 4.5 of the paper, our model achieves superior results.