Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation - Supplementary

1. Detailed Non Local Upsampling

In our approach we introduce the Non-Local Upsampling (NLU) module which is used in place of conventional upsampling operations which are based on local information only (bilinear, deconvolution). The idea of the NLU is to upsample the semantic features based on all the tokens coming from the skip connection by using a MSA block in the decoder.

The NLU module is detailed in Fig. 1. By using the same blocks as in [5], the skip connection is embedded into a query matrix $\boldsymbol{Q} \in \mathbb{R}^{(4N_p)\times C}$ while the keys and values are computed from the semantic low resolution features: $\boldsymbol{K} \in \mathbb{R}^{N_p \times C}$ and $\mathbf{V} \in \mathbb{R}^{N_p \times C}$. The resulting attention matrix is $\boldsymbol{A} \in \mathbb{R}^{(4N_p)\times N_p}$. To maintain the residual connection in the Transformer block, the low resolution features are upsampled and a linear projection adapts the number of channels before the sum. Then a Feed Forward (FF) layer is also used. It is worth noting that a normalization layer is included in both parts but omitted in the schema for clarity. At the end, a concatenation of the skip-connection and the upsampled semantic features ends the NLU the same way than in the standard U-Net architectures.

2. GLAM-Transformer complexity

The computational complexity of an MSA module for an image I divided into $h \times w$ patches has quadratic scaling with respect to the image area hw. The windowed approach W-MSA only depends on $N_p hw$. The complexity of both methods is given by:

$$\Omega(\text{MSA}(I)) = 4hwc^2 + 2(hw)^2 c \tag{1}$$

$$\Omega(\text{W-MSA}(I)) = 4hwc^2 + 2N_p hwc \tag{2}$$

This makes the W-MSA scalable to a large number of patches where the MSA can not be computed. With few global tokens, the global attention adds only a few numbers of operations as it corresponds to adding N_g tokens in each window and performing MSA over a sequence of length $N_g \times N_r$. It is also worth noting that the global tokens add a limited memory overhead as they do not require any more activation saving and only add a few elements in the attention matrix from each transformer block.

3. Detailed Experimental Settings

3.1. ADE20K and Cityscapes

For both ADE20K and Cityscapes, we implemented GLAM into the mmseg codebase [3]. All experiments ran on 8 Tesla V100 GPUs with 32GB and a batch size of 16 using data augmentation from the mmseg framework : random horizontal flipping, random re-scaling within ratio range [0.5, 2.0] and random photometric distortion. GLAM is implemented into the Swin and Swin-Unet models. Therefore, we were able to use the pre-trained weights from the respective models on ImageNet-1k [4]. For the case of the Swin-Unet backbone, we keep the same strategy as in [1] and duplicates symmetrically the encoder's weights to the decoder before fine-tuning. The added NLU and G-MSA modules could not benefit from this strong pre-training and their parameters were initialized randomly. Thanks to their integration into the overall architecture and the limited parameter increase they represent, this did not impact the good performances of the GLAM models. Complete pre-training on ImageNet of the GLAM backbones may however lead to even higher scores. The chosen optimizer is Adam with weigh decay of 0.01 and a polynomial learning rate scheduler starting from 0.00006 and with a factor of 1.0. The images in train are cropped at a size of 512×512 for ADE20K and 768×768 for Cityscapes. In validation the complete image is provided.

3.2. Synapse

Synapse is a medical image dataset composed of abdominal CT-scans. Thus, the models aren't pretrained on ImageNet as for ADE20K or Cityscapes. However, we integrated our experiments in the nnUnet framework that integrates an efficient training procedure. We follow the nnFormer model and used the SGD optimizer with an initial learning rate of 0.01. We employ a polynomial learning rate scheduler and a weight decay of 3e-5. The loss function is a combination of the cross entropy and dice. Similarly than nnFormer, the numbers of heads used in the en-

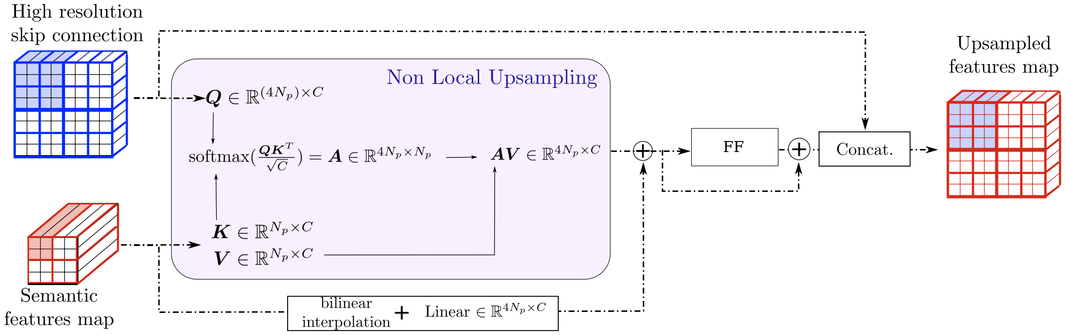

High resolution
skip connection
Non Local Upsampling
$\boldsymbol{Q} \in \mathbb{R}^{(4N_p)\times C}$
$\mathrm{softmax}(\frac{\boldsymbol{QK}^T}{\sqrt{C}}) = \boldsymbol{A} \in \mathbb{R}^{4N_p\times N_p}$
$\boldsymbol{AV} \in \mathbb{R}^{4N_p\times C}$
$\boldsymbol{K} \in \mathbb{R}^{N_p\times C}$
$\boldsymbol{V} \in \mathbb{R}^{N_p\times C}$
FF
Concat.
Upsampled
features map
Semantic
features map
bilinear
interpolation
+
Linear $\in \mathbb{R}^{4N_p\times C}$

Figure 1. **Non-Local Upsampling** The upsampling is processed window by window and is conceived as a super-resolution module where the low resolution feature map in the decoder (red) are re-embedded based on the high resolution ones coming from the encoder (blue). The patches are downsampled by a factor 2 before each hierarchy in the models. A given region from the decoder corresponds then to four neighbouring windows in the feature map coming from the skip connection.

coder stages are [6, 12, 24, 48]. The training is performed through 1000 epochs where each image is cropped at a size of $(128\times128\times64)$, as it is classically done for semantic segmentation over large 3D medical images, and in validation, we use a sliding window on the complete input volume.

4. Additional Results

ADE20K In this additional experiments we use Multi Scales (MS) inference to evaluate the model and their extended GLAM version on ADE20K. As shown in 1, while MS inference improves the performances for all the methods, the GLAM models still outperform their baselines. Indeed, in this configuration, GLAM-Swin-UNet Base reach +1.55% on ADE20K and is still +0.93% higher than Swin-UNet Base.

Cityscapes We provide the same analysis on Cityscapes and compare the performances of Sinw-Unet Tiny and GLAM-Swin-Unet Tiny with and without MS inference as reported in 2. Again, GLAM-Swin-Unet Tiny outperforms Swin-Unet Tiny by 1% mIoU when trained over 40k epochs using MS inference. Moreover, we also give complementary results to the ones reported in the Tables 1 and 2 of the main paper by providing performances with both models trained through 160k iterations. As can be seen in 2, the better performances of the GLAM model are stable as the GLAM-Swin-Unet outperfoms its baseline by 0.80% mIoU and 1.09% mIoU with respctively SS and MS inference when trained through 160k epochs.

Synapse To explore more in depth the performance gain brought out by GLAM in Table 3 of the main paper, we show in 3 the segmentation results for the different organs of the dataset. The results are given for two recent baselines : TransUNet [2] and nnFormer [6] as well as for GLAM-nnFormer. We use the publicly available implementations provided by authors for both models[1,2]. The proposed GLAM-nnFormer sensibly outperform both baselines for all the classes except on the kidneys and the pancreas where the result are close to the standard nnFormer.

Table 1. **GLAM Improvements with Multi Scale inference on ADE20K.** Performances are evaluated with respect to mIoU for single scale inference (SS) and multiscales inference (MS).

Method	Size	SS	MS
Swin-Unet [1]	Tiny	42.75	44.72
GLAM-Swin-Unet	Tiny	**44.19**	**46.11**
Swin-Unet [1]	Base	47.85	49.72
GLAM-Swin-Unet	Base	**49.10**	**50.65**

Table 2. **GLAM Improvements with Multi Scale inference on Cityscapes.** Performances are evaluated with respect to mIoU for single scale inference (SS) and multiscales inference (MS).

Method	Size	SS	MS
Swin-Unet 40K [1]	Tiny	77.43	78.56
GLAM-Swin-Unet 40K	Tiny	**78.29**	**79.56**
Swin-Unet 160K [1]	Tiny	79.98	80.90
GLAM-Swin-Unet 160K	Tiny	**80.78**	**81.99**

Global token merging strategy. Here, we study the importance of how the global tokens between different windows are merged: with the GLAM transformer, we use a global self-attention (G-MSA) mechanism. We compare G-MSA with an averaging and a random permutation strategy. We can see in Table 4 that G-MSA is largely superior to the two other options. This validates the usefulness of the G-MSA step, which enable to indirectly model full range interactions between visual region when applied after W-MSA.

5. Visualizations

ADE20K and Cityscapes In 3 and 4, we select some representative images of the GLAM-Swin-Unet and

[1] https://github.com/Beckschen/TransUNet
[2] https://github.com/282857341/nnFormer

Table 3. Detailed per-organ comparison on the multi-organ Synapse dataset (Dice Score in %).

Methods	Aotra	Gallbladder	Kidnery(L)	Kidnery(R)	Liver	Pancreas	Spleen	Stomach
TransUNet [2]	87.23	63.16	81.87	77.02	94.08	55.86	85.08	75.62
nnFormer [6]	89.81	63.18	93.78	**94.58**	96.19	**83.16**	95.76	86.14
GLAM-nnFormer	**90.10**	**65.81**	**93.92**	**94.56**	**96.74**	82.91	**96.49**	**88.20**

Table 4. Global token merging (tiny Swin-Unet, ADE20k).

Merging strategy	mIoU
Random permutation	43.2
Average	43.7
G-MSA Merging	**44.2**

GLAM-Swin-Upernet. We provide attention maps for the lowest hierarchy as well as the generated segmentation map for the GLAM models. The attention is computed with respect to a global token associated to the 7×7 blue window plotted in the image (not to the scale). For the first stage of the model, the patch size is 4×4 patches and thus the dimension of the window is 28×28 pixels. We see that the model manages to detect long range interactions directly in high resolution features map without being limited by the small window size. Attention is paid mostly between elements of the same class : vegetation in 3, chairs or sky in 4 but also to salient elements such as corner or edges and semantic ones such as cars and pedestrians.

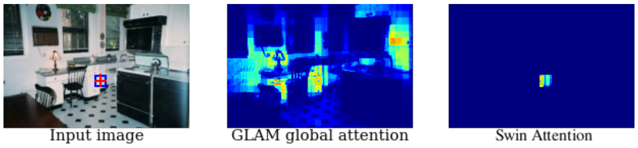

Input image
GLAM global attention
Swin Attention

Figure 2. Global attention of GLAM compared to vanilla Swin on ADE20K.

We provide another comparison on ADE20K in Fig. 2 below. Again, we can notice that Swin's attention is limited to the small blue region. In contrast, GLAM can compute a global attention map at high resolution thanks the G-MSA module, providing both accurate spatial information and global context.

Synapse In 5, we present more segmentation results on 3D medical images and provide a qualitative analysis of the performances between nnFormer and GLAM-nnFormer. The GLAM model manages to retrieve better segmentation of the liver (pink) and the stomach (purple). The memory effect of the global tokens manage to limit the error due to the inference on 3D crops which is well illustrated on the liver reconstruction.

References

[1] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021.

[2] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[3] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[6] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation, 2021.

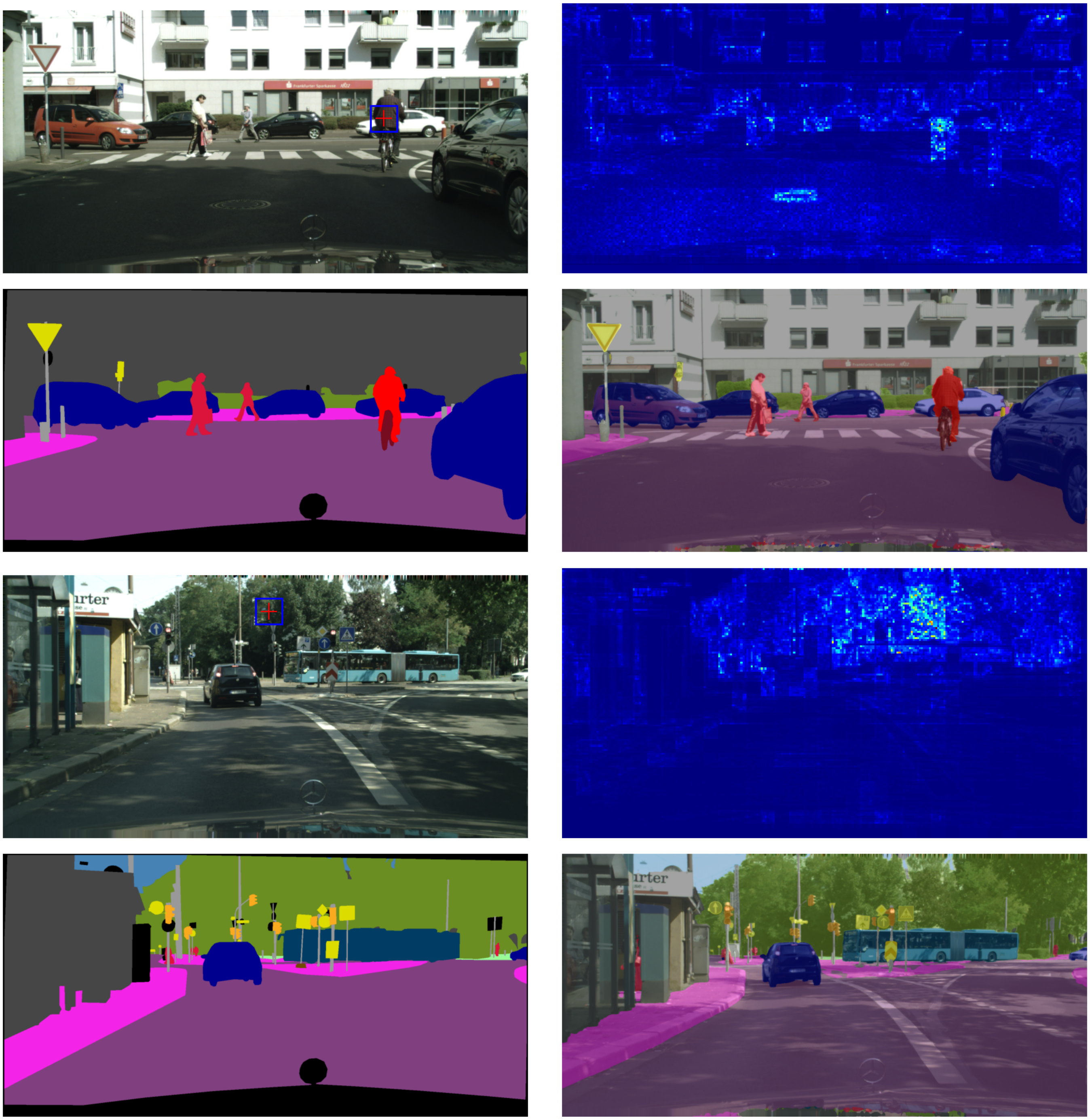


Figure 3. **Qualitative results of GLAM incorporated to Swin-upernet on Cityscapes** For each scene, we show from top-left to bottom-right : the image, the global attention map with respect to the blue window, the ground truth and the predicted segmentation.

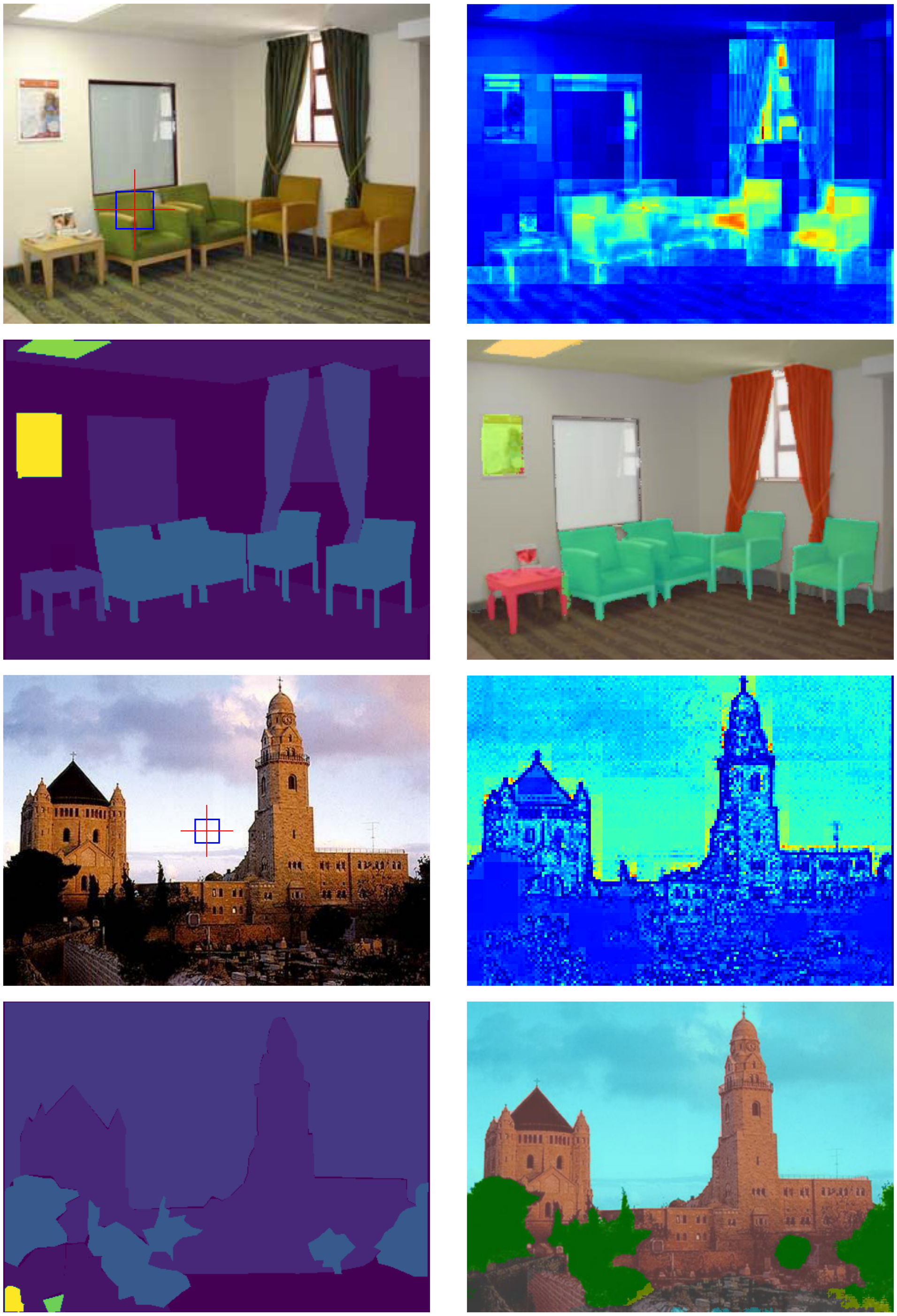
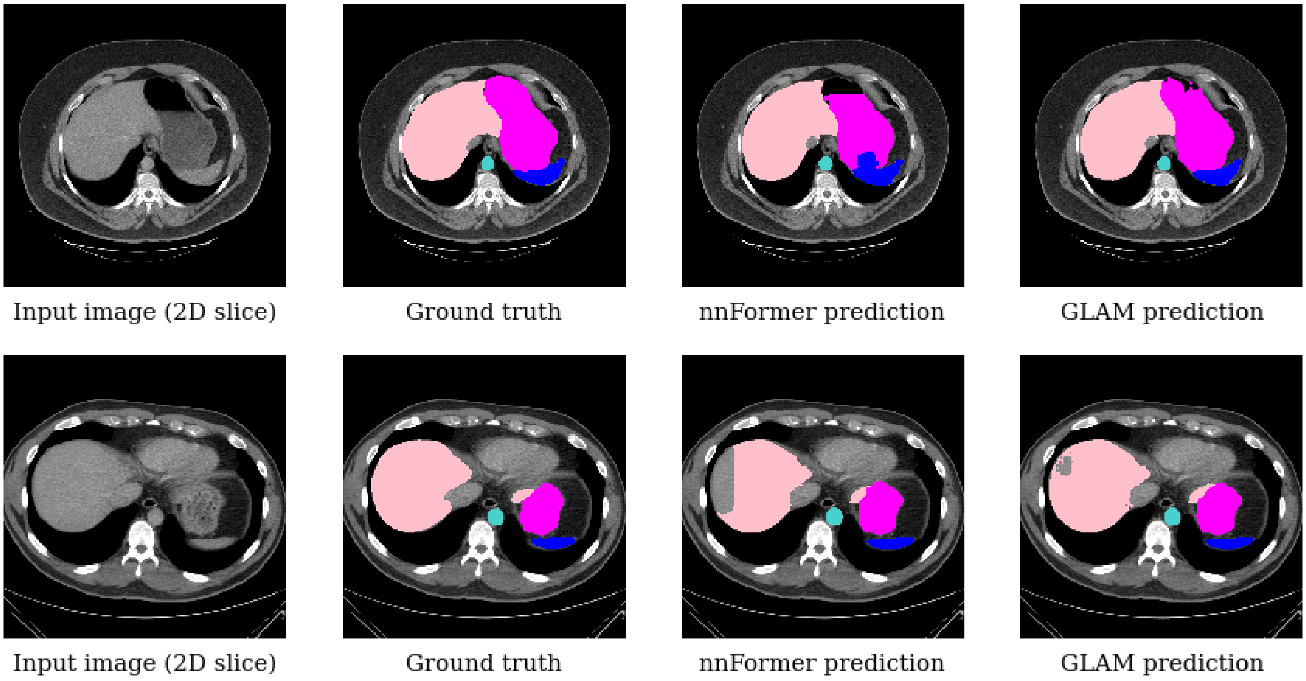


Figure 4. **Qualitative results of GLAM incorporated to Swin-UNet on ADE20K** For each scene, we show from top-left to bottom-right : the image, the global attention map with respect to the blue window, the ground truth and the predicted segmentation.



Input image (2D slice)

Ground truth

nnFormer prediction

GLAM prediction

Input image (2D slice)

Ground truth

nnFormer prediction

GLAM prediction

Figure 5. **Qualitative results of GLAM Incorporated to nnFormer on Synapse.** The obeserved organs are the liver (pink), the stomach (purple), the aorta (cyan) and the spleen (blue).