Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation

Loïc Themyr, Clément Rambour, Nicolas Thome, Toby Collins, Alexandre Hostettler

Full Contextual Attention for Multi-resolution Transformers in Semantic Segmentation

Loic Themyr[1,2] Clement Rambour[1] Nicolas Thome[1,3] Toby Collins[2]
Alexandre Hostettler[2]
[1]Conservatoire National des Arts et Métiers, Paris, France
[2]IRCAD, Strasbourg, France
[3]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
loic.themyr@lecnam.net

Abstract

Transformers have proved to be very effective for visual recognition tasks. In particular, vision transformers construct compressed global representations through self-attention and learnable class tokens. Multi-resolution transformers have shown recent successes in semantic segmentation but can only capture local interactions in high-resolution feature maps. This paper extends the notion of global tokens to build GLobal Attention Multi-resolution (GLAM) transformers. GLAM is a generic module that can be integrated into most existing transformer backbones. GLAM includes learnable global tokens, which unlike previous methods can model interactions between all image regions, and extracts powerful representations during training. Extensive experiments show that GLAM-Swin or GLAM-Swin-UNet exhibit substantially better performances than their vanilla counterparts on ADE20K and Cityscapes. Moreover, GLAM can be used to segment large 3D medical images, and GLAM-nnFormer achieves new state-of-the-art performance on the BCV dataset.

1. Introduction

Transformers have achieved state-of-the-art performances in various Natural Language Processing (NLP) tasks [35]. Recently, fully transformer-based models have reached excellent performances on vision tasks such as image classification [12] and semantic segmentation [46].

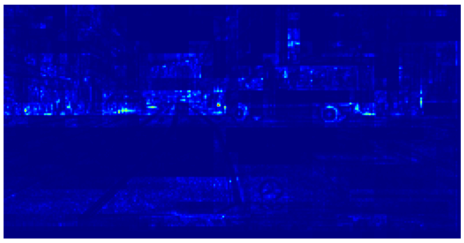
The main appeal of transformers is their ability to grasp long-range interactions, which is a crucial point for semantic segmentation. However, this strategy is not easily scalable to high-resolution images involving a large number of patches, due to the quadratic complexity of the transformer's attention module. A simple and efficient strategy to tackle this limitation is to rely on multi-resolution approaches, where the attention in high-resolution feature maps is computed on sub-windows. There have been various recent attempts in this direction [24, 38, 44, 37, 2]. However, they limit the interactions of high-resolution features to within each window.

We introduce an approach for semantic segmentation that incorporates global attention in multi-resolution transformers (GLAM). The GLAM module enables full-range interactions to be modeled at all scales of a multi-resolution transformer. As illustrated in Fig. 1, incorporating GLAM into the Swin architecture [24] enables to jointly capture fine-grained spatial information in high-resolution feature maps and global context, where both elements are crucial for proper segmentation in complex scenes. This concept is illustrated in Fig. 1 where Fig. 1a) shows an input image, and Fig. 1b) shows the self-attention map provided by GLAM in the highest-resolution feature map for the pedestrian region pointed out by the yellow cross in Fig. 1a). We can see that the attention map involves long-range interactions between other visual structures (cars, buildings), in contrast to the Swin baseline, where the window attention at a high-resolution feature map is limited to the small rectangular region in Fig. 1a). Consequently, GLAM has exploited longer-range interactions to successfully segment the image, as shown in 1d).

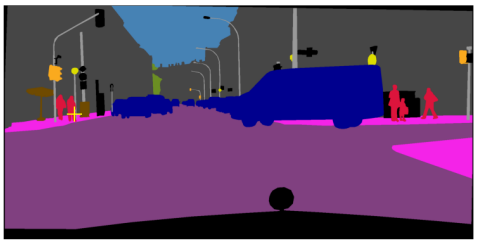
To achieve this goal, we have made the following novel contributions:

- We introduce the GLAM transformer, able to represent full-range interactions between all local features at all resolution levels. The GLAM transformer is based on learnable global tokens interacting between all visual features. To fully take into account the global context, we also design a non-local upsampling scheme (NLU).
- GLAM is a generic module that can be incorporated into any multi-resolution transformer. It consists of a

a) Input image (size 1025x2049)



b) GLAM global attention



c) Ground Truth

d) GLAM prediction

Figure 1. When segmenting the high-resolution image in a) with state-of-the-art multi-resolution transformers, *e.g*. Swin [24], the attention in the highest-resolution feature maps is limited to a small spatial region, *i.e*. the blue square for the yellow-crossed pedestrian. Our method incorporates GLobal Attention in Multi-resolution transformers (GLAM). The GLAM attention map for the pedestrian in a) is depicted in b): it captures both fine-grained spatial information and long-range interactions, enabling successful segmentation, as shown in d).

succession of two transformers applied on the merged sequence of global and visual tokens and in-between global tokens. We highlight that the GLAM transformer can represent full-range interactions between image regions at all scales while retaining memory and computational efficiency. Beyond spatial interactions, global tokens also model the expected scene composition.

- Experiments on various generic (ADE20K), autonomous driving (Cityscape) and medical (Synapse) datasets show the important and systematic gain brought by GLAM when included into existing state-of-the-art multi-resolution transformers including Swin, Swin-Unet, and nn-Former. We also show that GLAM outperforms state-of-the-art methods on Synapse. Finally, ablation studies, model analysis, and visualizations are presented to assess the behavior of GLAM.

2. Related work

Semantic Segmentation. In the deep learning era, Fully Convolutional neural Networks (FCNs) [25, 33, 45, 6, 40] have mainly led state-of-the-art performance in semantic segmentation. For example, DeepLab [6] is a model based on an encoder-decoder architecture, and U-shape networks [32] and 3D variants [26, 18] are extremely popular in medical image segmentation. However, those models are limited to a local receptive field which is small for high-resolution images. Recently, transformers [35] have gained a lot of interest from their ability to model long-range interactions, which allows larger spatial context information to be exploited.

Vision Transformer backbones. Building on the strong performances of transformers for auto-regressive inference, fully transformer-based models for image generation have been proposed [7]. Other early works proposed models combining CNNs and attention for vision tasks such as object detection [49], disparity estimation [23] or semantic segmentation [16]. More recently, fully transformer architectures have outperformed FCN baselines in image classification. ViT [12] is the first pure transformers backbone that achieved state-of-the-art performance for image classification but it requires very large training databases. DeiT [34] managed to reduce this requirement through data-efficient training strategies and distillation.

Multi-resolution transformers. Several recent approaches proposed adaptations of the vanilla ViT architecture. In particular, some architectures rely on multi-resolution processing. T2T ViT [42] constructs richer semantic feature map through token aggregation while TnT [15] and crossViT [4] uses two transformers for fine and coarse resolution. PvT [38] is the first backbone with a fully pyramidal architecture that is based on windowed transformers, allowing to process the images at fine resolution and to build rich feature maps, while reducing the spatial complexity. Other methods kept this hierarchical approach while improving information sharing between the windows. Swin [24] and its variant [2, 48] proposed to used shifted windows, Twins [9] uses interleaved fine and coarse resolu-

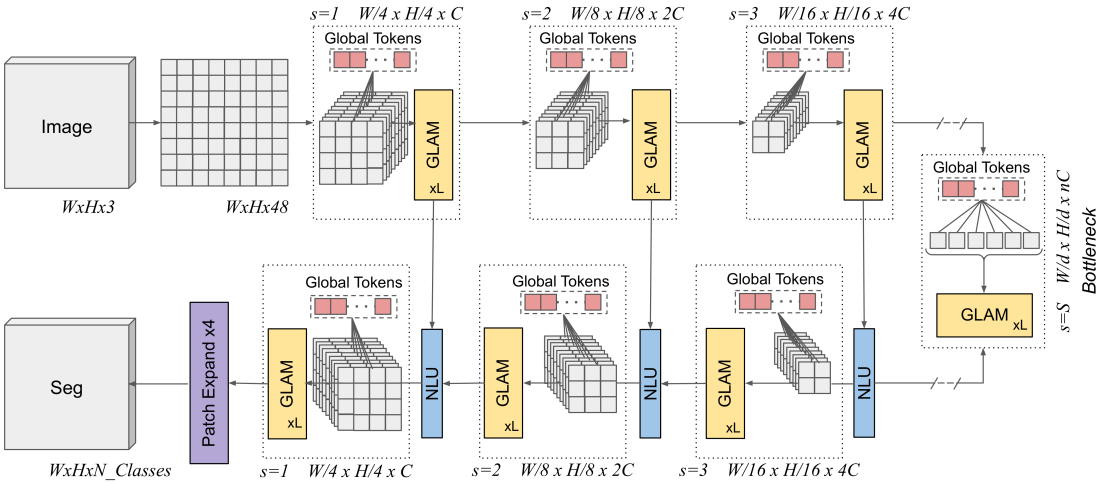


Figure 2. The GLAM module for modeling full-range interaction in multi-resolution transformers. GLAM is included at each resolution level of any multi-resolution transformer architecture, *e.g*. Swin-Unet [24] or Swin-UperNet [24]. GLAM includes learnable global tokens, which are leveraged into a succession of two attention steps. We show that this design can indirectly represent long-range interactions between all image regions at all scales, and also external information useful for segmentation while retaining efficiency. We also introduce a non-local upsampling scheme (NLU) to extend the global context modeling in full transformer U-shape architectures such as [2, 48].

tion transformers, and CvT [39] replaces linear embedding with convolutions.

Efficient Self-Attention. Long sequences have been a challenge for transformers because the original self-attention mechanism has a quadratic complexity in the sequence length. To tackle this, many approaches focus on designing efficient self-attention mechanisms.

Most of them are developed for NLP tasks and can be grouped into four categories. The first category uses a sparse approximation of the attention matrix [30, 20, 28]. Among these approaches, window-based patch extraction vision transformers recently provided a simple yet efficient approach to compute attention [24, 37, 13]. The second category is composed of methods based on a low-rank approximation of the attention matrix, such as Linformer [36]. The third category (memory-based transformers) construct buffers of extra tokens used as static memory [31, 22]. The fourth category (kernel-based methods) provides a linear approximation of the softmax kernel [8, 29, 19]. Some vision transformers have combined multiple efficient attention mechanisms. The recent ViT-inspired backbone PvT [38] is based on windowed self-attention and attention approximation close to Linformer. ViL [43] balances sparse attention by using a reduced set of global tokens (usually a single one) to extract global representations of the input image.

The GLAM method introduced in this paper is a window- and memory-based transformer that fits well in existing multi-resolution vision backbones. Unlike most other multi-resolution backbones, GLAM fully extends the notion of windowed attention to semantic segmentation by introducing global tokens at the window level, and by designing a specific GLAM transformer cascading window (W-MSA) and global (G-MSA) attention. We highlight that GLAM enables global communication across all image regions and also encodes learned information from all the training sets.

3. The GLAM Method

The main idea in GLAM is to provide a way to represent full range interactions at all feature map resolutions, which is impossible in vanilla models, especially in high-resolution feature maps, due to the quadratic complexity of attention transformers.

GLAM is illustrated in Fig. 2, where it has been added to the Swin-Unet architecture [24]. Note that GLAM can be included in various multi-resolution architectures, *e.g*. Swin [2] or PvT [38] and is also applicable for 3D segmentation, *e.g*. nn-Former [48]. The core idea in GLAM is to design global tokens (in red in Fig. 2), which are leveraged into a succession of two attention steps: first, between visual tokens in each window independently and, second, between global tokens among different windows. We show in Fig. 3.2 that this design enables to represent full range interactions between all image regions at all scales, and also external information useful for segmentation, while retaining efficiency. We also introduce a non-local upsampling scheme (NLU) to extend the full context modeling in U-shape architectures and to provide an efficient interpolation of rich semantic feature maps in an associated decoder.

3.1. Multi-resolution transformer architecture

As shown in Fig. 2, GLAM can be included into any multi-resolution transformer architecture [24, 38, 44, 37, 2, 48].

Transformer. At each resolution level s, given a sequence of visual tokens, a transformer learns representations through Self-Attention (SA). SA is given by the expectation of each token with respect to their probability of sharing the same embedding. The Multi-head Self-Attention (MSA) is obtained from the linear combination of m parallel SA operations. Finally, a complete transformer module is obtained by plugging the output of the MSA into a Multi-Layer Perceptron (MLP). Layer norm operations, as well as residual connections, are added respectively before and after MSA and MLP modules.

Windowed attention. MSA cannot be applied to long sequences *e.g.* patches from high-resolution images becasue the computation of the attention matrix has quadratic memory complexity. To allow high-resolution processing and thus long sequences of small patches, windowed transformers treat the image as a batch of non-overlapping windows [24, 38, 37, 44]. This approach is combined with a pooling strategy [2, 24, 38, 44] and is well suited to build a multi-resolution encoder, able to produce rich semantic maps. Multi-resolution backbones are built by chaining windowed transformer blocks and downsampling. These hierarchical architectures manage to build larger receptive fields in deeper layers, similarly to CNNs. This, however, does not guarantee a global receptive field and the maximal receptive field depends on the model's depth. More importantly, this processing introduces a major modification to the transformer modules. At a finer resolution, only local interactions are considered. With this modification, the processing of isolated patches by self-attention may not be as effective as global self-attention performed on the full image.

3.2. Global attention multi-resolution transformers

We show how the GLAM module can provide global attention in all feature maps of multi-resolution transformers. The GLAM transformer is illustrated in Fig. 3, consisting in a sequence of L transformer blocks, processing visual tokens in each region of the multi-resolution maps (shown in blue in Fig. 3) and global tokens (shown in red in Fig. 3).

The basic idea behind GLAM is to associate global tokens to each window that are responsible to encapsulate the local information and transmit it to other image regions by computing MSA between all global tokens. Thus, when information is processed at the window scale, the visual tokens embedding incorporate useful long-range information.

Global Tokens. Global tokens lie at the core of Global Attention (GA). They are specific tokens concatenated to each window and are responsible for communication between windows. We define as N_r the number of windows in the feature map, N_p as the number of patches per window, and $\{\boldsymbol{w}_k^l\}_{1\leq k\leq N_r}$ as the sequence of windows after being processed by the l^{th} GLAM-transformer block. We define as $\{\boldsymbol{g}_k^l\}_{1\leq k\leq N_r}$ the sequence of N_g-dimensional global tokens associated to each window. The initialization of the global tokens $\{\boldsymbol{g}_k^0\}_{1\leq k\leq N_r}$ is the same for all windows and is learned by the model. The input of the l^{th} transformer block, defined as $\boldsymbol{z}^l$, is a batch of tokens from each window concatenated with the corresponding global tokens, *i.e.* $\boldsymbol{z}^l \in \mathbb{R}^{N_r\times(N_g+N_p)\times C}$ with C being the dimension of the tokens. Consequently, the elements in the batch have the form:

$$\forall k \in [1..N_r], \boldsymbol{z}_k^l = \begin{bmatrix} \boldsymbol{g}_k^l \\ \boldsymbol{w}_k^l \end{bmatrix} \in \mathbb{R}^{(N_g+N_p)\times C}. \tag{1}$$

GLAM-Transformer. The communication between windows at a given hierarchy level is obtained through the interaction of global tokens. At each block l of the GLAM-transformer, there are two steps: *i)* visual tokens grasp their local statistics through a local window transformer (W-MSA), and *ii)* the global tokens are re-embedded by a global transformer (G-MSA), where global tokens from different windows interact with each other. Formally, the l^{th} GLAM-transformer block inputs $\boldsymbol{z}^{l-1}$ and outputs $\boldsymbol{z}^l$ by the succession of a W-MSA and a G-MSA step:

$$\begin{aligned} \hat{\boldsymbol{z}}^l &= \text{W-MSA}(\boldsymbol{z}^{l-1})\,, \\ \boldsymbol{g}^l &= \text{G-MSA}(\hat{\boldsymbol{g}}^l)\,, \\ \boldsymbol{z}^l &= \left[\boldsymbol{g}_k^{l^T}\, \hat{\boldsymbol{w}}_k^{l^T}\right]^T \end{aligned} \tag{2}$$

We define as $\boldsymbol{A}_r^l$ the attention matrix for the window r in the transformer block l. We introduce the following decomposition to express the attention with respect to the global and local tokens:

$$\boldsymbol{A}_r^l = \begin{bmatrix} \boldsymbol{A}_{r,gg}^l & \boldsymbol{A}_{r,gw}^l \\ \boldsymbol{A}_{r,wg}^l & \boldsymbol{A}_{r,ww}^l \end{bmatrix}. \tag{3}$$

The square matrices $\boldsymbol{A}_{r,gg}^l \in \mathbb{R}^{N_g\times N_g}$ and $\boldsymbol{A}_{r,ww}^l \in \mathbb{R}^{N_p\times N_p}$ give the attention from the global token and the spatial tokens on themselves respectively. The matrices $\boldsymbol{A}_{r,gw}^l \in \mathbb{R}^{N_g\times N_p}$ and $\boldsymbol{A}_{r,wg}^l \in \mathbb{R}^{N_p\times N_g}$ are the cross attention matrices between local and global tokens. We define as $\boldsymbol{B}^l \in \mathbb{R}^{(N_r\cdot N_g)\times(N_r\cdot N_g)}$ the global attention matrix from all the global token sequence and $\boldsymbol{B}_{ij}^l \in \mathbb{R}^{N_g\times N_g}$ as the sub-matrices giving the attention between the global tokens of windows i and j.

GLAM-Transformer properties. Putting aside the value matrix, the W-MSA gives the following embedding $\hat{\boldsymbol{g}}_r^l$ from $\boldsymbol{g}_r^{l-1}$:

$$\hat{\boldsymbol{g}}_r = \boldsymbol{A}_{r,gg}^l \boldsymbol{g}_r^{l-1} + \boldsymbol{A}_{r,gw}^l \boldsymbol{w}_r^{l-1}\,. \tag{4}$$

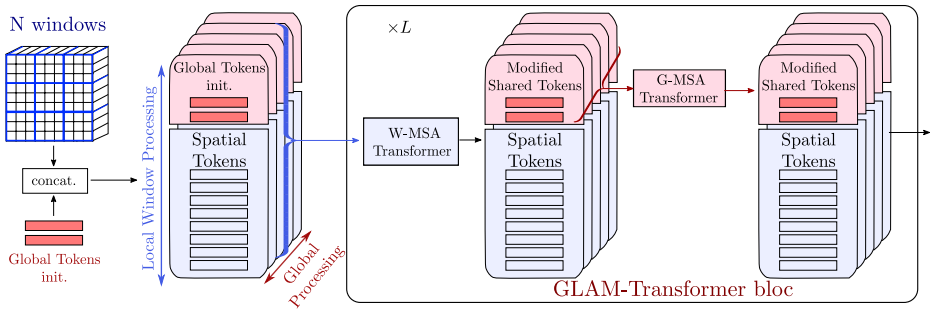


Figure 3. **GLAM-Transformer:** as in multi-resolution approaches, each input feature map is divided into N_r non overlapping windows (blue). The core idea in GLAM is to design learnable global tokens (in red). The visual tokens from each window are concatenated with the global tokens and processed through a local window transformer (W-MSA). Every W-MSA is followed by a global transformer (G-MSA), where global tokens between different windows interact with each other, which brings a global representation to each window. These two steps give the GLAM-Transformer block; Multiple blocks are chained at every hierarchy level in typical multi-resolution transformer backbones. We show that global tokens learned from GLAM-Transformer indirectly model global interactions between all visual tokens in all widows. The global tokens are also able to represent extra learnable knowledge beyond the patch interactions in a single image.

The G-MSA, *i.e.* the MSA on the sequence of global tokens gives the following embeddings:

$$\begin{aligned} \boldsymbol{g}_r^l &= \sum_{n=1}^{N_r} \boldsymbol{B}_{rn}^l \hat{\boldsymbol{g}}_n^l \\ &= \sum_{n=1}^{N_r} \boldsymbol{B}_{rn}^l (\boldsymbol{A}_{r,gg}^l \boldsymbol{g}_r^{l-1} + \boldsymbol{A}_{r,gw}^l \boldsymbol{w}_r^{l-1}) \,. \end{aligned} \tag{5}$$

From Eq. (5), we have the expression of the global token for a window r processed by the l G-MSA block transformer. Developing this formulation we obtain the following expression for the k^{th} global token in the r^{th} window:

$$\begin{aligned} g_{k,r}^l &= \sum_{r'=1}^{N_r} \sum_{j=1}^{N_g} b_{k,r,j,r'} \left(\sum_{i=1}^{N_g+N_p} a_{j,r',i} z_{i,r'}^{l-1} \right) \\ &= \sum_{r'=1}^{N_r} \sum_{j=1}^{N_g} b_{k,r,j,r'} \left(\sum_{i=1}^{N_g} a_{j,r',i} g_{i,r'}^{l-1} \right. \\ &\quad \left. + \sum_{i=1}^{N_p} a_{j,r',i+N_g} w_{i,r'}^{l-1} \right). \end{aligned} \tag{6}$$

The variables $z_{i,r}$, $g_{i,r}$ and $w_{i,r}$ corresponds respectively to the visual, global or generic token i in window r. $a_{j,r,i}$ is the attention coefficient given by the token j to the token i inside the window r. $b_{j,r,i,r'}$ is the attention coefficient from the global token j in the window r to the global token j in the window r'. Re-arranging the indices of equation Eq. (6) leads to the following expression for the k^{th} global token in the r^{th} window:

$$\begin{aligned} g_{k,r}^l &= \sum_{r'=1}^{N_r} \sum_{i=1}^{N_p} \left(\sum_{j=1}^{N_g} b_{k,r,j,r'} \; a_{j,r',(i+N_g)} \; w_{i,r'}^{l-1} \right) \\ &\quad + \sum_{r'=1}^{N_r} \left(\sum_{j=1}^{N_g} b_{k,r,j,r'} \sum_{i=1}^{N_g} a_{j,r',i} \; g_{i,r'}^{l-1} \right) \end{aligned} \tag{7}$$

This leads to a global attention matrix $\boldsymbol{G}_k \in \mathbb{R}^{(N_r \cdot N_p) \times (N_r \cdot N_p)}$ associated to the k^{th} global token given by $[\boldsymbol{G}_k]_{r',i} = \sum_{j=1}^{N_g} b_{k,r,j,r'} a_{j,r',(i+N_g)} + \sum_{j=1}^{N_g} b_{k,r,j,r'} \sum_{i=1}^{N_g} a_{j,r',i}$. Eq. (7) gives the embedding of the the global token $g_{k,r}^l$ at the l^{st} GLAM-transformer block, with respect to all visual tokens in all feature map windows $w_{i,r'}^{l-1}$ (first row), and all global tokens $g_{i,r'}^{l-1}$ (second row). This rewriting shows that the global embedding $g_{k,r}^l$ captures interactions between all image regions independently of the resolution. The different terms in the decomposition are interpreted as an attention map associated with each image region. This is the visualization shown in Fig. 1: the row of the first term corresponds to patch-based attention which depends on all the tokens of the feature map, while the second row represents window-based attention.

Overall, global tokens embedded with GLAM-transformers provide a way for information propagation across all windows (first row in Fig. 7), but also global information (second row) that goes beyond matching visual features in a single image. Especially, this represents global and learned information across the dataset, and can be leveraged as a stabilizing effect in SA, because the information is shared not only from the input but from all the windows in the dataset. This makes them a powerful tool to interpret isolated tokens and to take advantage of redundant structures in the data.

Non-Local Upsampling. We introduce a Non-Local Upsampling (NLU) module for a fully transformer decoder such as [2, 48]. NLU is designed to upsample the semantic features based on all the tokens coming from the skip connection, by drawing inspiration for non-local means [1].

The proposed NLU is illustrated in the supplementary material. To perform the upsampling, the skip connections are embedded into a query matrix of size $(4N_p) \times C$ while the semantic low-resolution features are embedded into the keys and values of size $N_p \times C$. The projection of the values on the resulting attention matrix has the size $(4N_p) \times C$.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluated on three different semantic segmentation datasets: ADE20K [47], Cityscapes [11] and Synapse [21]. ADE20K is a scene parsing dataset composed of 20,210 images with 150 object classes. Cityscapes contains driving scenes and is composed of 5,000 images annotated with 19 different classes. Synapse is an abdominal organ segmentation dataset that includes 30 Computerized Tomography (CT) scans which are 3D volumes annotated with 8 abdominal organs.

Implementation details. GLAM models were implemented into the mmseg [10] codebase and the models were trained on 8 Tesla V100 GPUs. The layers were pretrained on ImageNet-1K and standard augmentation was used: random crop, rotations, translations, *etc.* More details are provided in supplementary. We used the Adam optimizer with a weigh decay of 0.01 and a polynomial learning rate scheduler starting from 0.00006 and with a factor of 1.0. The reported segmentation performances are mean Intersection over Union (mIoU) for ADE20k and Cityscapes and Dice Similarity Score (DSC) for Synapse.

4.2. GLAM performance

GLAM in multi-resolution transformers. GLAM is well suited to work with window transformers such as PvT [38, 37] or Swin [24] as well as its variants [2, 48]. Due to the top performances of Swin, we incorporated GLAM into this backbone to compute the segmentation of 2D datasets leading to two models: GLAM-Swin-UperNet and GLAM-Swin-Unet. The first one is a hybrid model combining a transformer backbone and a CNN head [2, 40] while the second one is a full transformer model with a decoder symmetric to the encoder [2]. For 3D images, GLAM was plugged into nnFormer [48] which is designed similarly to Swin-Unet for 3D medical image segmentation. The performances of the Swin and GLAM models are presented in Table 1. GLAM models exhibit important and consistent performance gains compared to their vanilla counterparts, either on small or larger models: *e.g.* $\sim$ +1.5pt gain on ADE20K with Swin-Unet (Base or Tiny), and +1.2pt on Synapse on the recent nn-Former model.

Table 1. **GLAM Improvements on various multi-resolution transformers.** Performances are evaluated with respect to mIoU for ADE20k and Cityscapes and average DSC for Synapse.

Dataset	Method	Size	Score
ADE20K	Swin-Unet [2]	Tiny	42.75
	GLAM-Swin-Unet	Tiny	**44.19**
	Swin-UNet [2]	Small	47.49
	GLAM-Swin-UNet	Small	**47.90**
	Swin-Unet [2]	Base	47.85
	GLAM-Swin-Unet	Base	**49.10**
	Swin-UperNet[24]	Tiny	43.69
	GLAM-Swin-UperNet	Tiny	**44.16**
	Swin-UperNet [24]	Small	47.72
	GLAM-Swin-UperNet	Small	**47.75**
	Swin-UperNet [24]	Base	47.99
	GLAM-Swin-UperNet	Base	**48.44**
Cityscapes	Swin-UperNet [24]	Tiny	78.24
	GLAM-Swin-UperNet	Tiny	**78.64**
	Swin-UperNet [24]	Base	80.79
	GLAM-Swin-UperNet	Base	**81.47**
	Swin-Unet [2]	Tiny	77.43
	GLAM-Swin-Unet	Tiny	**78.29**
Synapse	nnFormer [48]	Tiny	87.40
	GLAM-nnFormer	Tiny	**88.60**

State-of-the-art comparison. We now compare the GLAM-Swin models with existing approaches on the ADE20K [47], Cityscapes [11] and Synapse [21].

ADE20K and Cityscapes. Table 2 summarizes our results. To be fair, we compared models up to $\sim$ 150M parameters, and we report the top performances from the mmseg [10] benchmark for all methods, with 160K training epochs for all methods. Moreover, we compared only methods trained on 768×768 resolution images on Cityscapes. In this setup, GLAM-Swin-Unet yields 49.10% mIoU on ADE20K outperforming its vanilla Swin counterpart with at least 1.10% mIoU. GLAM-Swin-UperNet achieves 81.47 % mIoU on Cityscapes which is 1.58 % better than its Swin-Upernet counterpart.

Synapse. Table 3 reports our results and recent baselines for 3D medical segmentation. GLAM-nnFormer significantly outperforms all other existing methods by at least 1.2% average Dice. To the best of our knowledge, GLAM-nnFormer outperforms state-of-the-art on the Synapse dataset.

Table 2. Comparison to state of the art methods on ADE20K and Cityscapes. All experiments are made or reported are with single-scale inference.

Method	Backone	ADE20K mIoU	Cityscapes mIoU
FCN [33]	ResNet-101	41.40	77.34
CCNet [17]	ResNet-101	43.71	79.45
DANet [14]	ResNet-101	43.64	80.47
UperNet [40]	ResNet-101	43.82	80.10
DNL [41]	ResNet-101	44.25	79.41
PSPNet [45]	ResNet-101	44.39	79.08
DeepLabV3+ [6]	ResNet-101	45.47	79.41
Trans2Seg [38]	PVT-S	42.60	-
FPN [38]	PVT-L	42.10	-
TNT [15]	TNT-S	43.60	-
SETR-PUP [46]	DeiT-L	46.34	79.21
Swin-Unet [2]	Swin-B	47.85	-
Swin-UperNet [24]	Swin-B	47.99	80.79
Twins-SVT-L [9]	Twins-SVT	48.80	-
GLAM-Swin-Unet	Swin-B	**49.10**	-
GLAM-Swin-UperNet	Swin-B	48.44	**81.47**

Table 3. Comparison to state of the art methods on Synapse.

Methods	Average Dice Score (%)
VNet [26]	68.81
U-Net [32]	76.85
Att-UNet [27]	77.77
R50-Deeplabv3+ [6]	75.73
TransUNet [5]	77.48
Swin-Unet [2]	79.13
TransClaw U-Net [3]	78.09
nnUNet (3D) [18]	86.99
nnFormer [48]	87.40
GLAM-nnFormer	**88.60**

4.3. Model Analysis

In this part, we analyze various important aspects of GLAM.

Number of Global Tokens. The number of global tokens directly influences the capacity of GLAM to model global interactions between the windows. Fig. 4 shows the impact of this hyper-parameter on segmentation performances. We can see that using more global tokens improves performances. However, it also increases the number of parameters and memory cost which forces a trade-off. We keep a reasonable value of 10 global tokens, which gives an important performance boost of +1.4pts in both the tiny and base versions of the Swin-Unet model.

Impact of NLU. GLAM improves context modeling in multi-resolution transformers thanks to global attention and Non-Local Upsampling (NLU). Table 4 provides an ablation study of these two components. We can see that NLU gives an improvement 0.45pt compared to the original Swin-Unet that uses a patch expension operation for upsampling. GLAM brings another large improvement for a total gain of +1.44pts compared to the baseline.

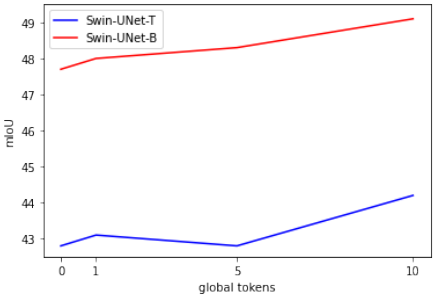


Figure 4. Impact of the number of global tokens on performance (mIoU) using ADE20k.

Table 4. Impact of the NLU and the GLAM transformer on a tiny Swin-Unet, 10 global tokens, on ADE20k.

Method	NLU	GLAM	mIoU
Swin-Unet-T			42.75
Swin-Unet-T	✓		43.20
Swin-Unet-T	✓	✓	**44.20**

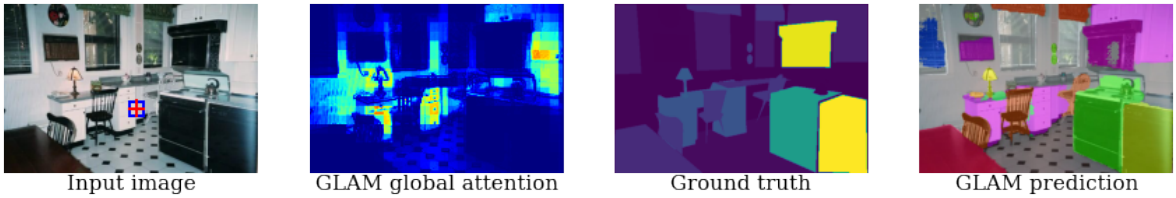
Long-range interaction. To highlight the impact of G-MSA, Table 5 shows the performances of GLAM backbones using only a W-MSA step but no G-MSA. GLAM backbones show consistent gains compared to their counterparts without G-MSA. This ablation highlight the crucial role of this step to leverage long-range interactions and that the performance gains made by GLAM can not only be explained by the parameter overhead.

Table 5. Impact of G-MSA phase on GLAM transformer on different model, 10 global tokens, on ADE20k. GLAM-nogmsa is GLAM without the G-MSA phase.

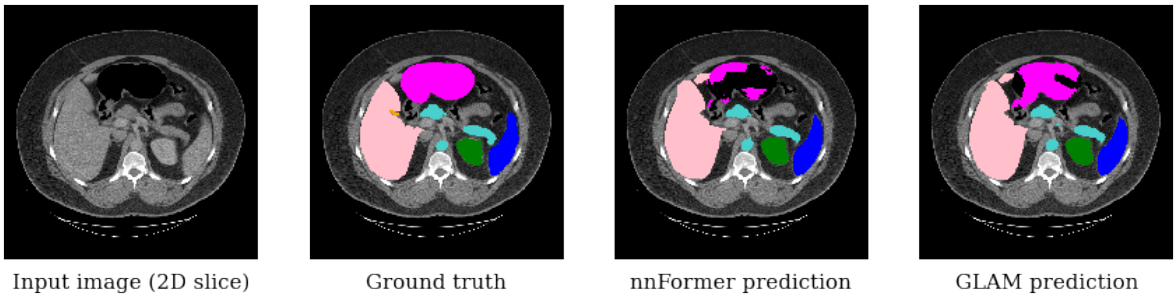
Method	mIoU
GLAM-nogmsa-Swin-Unet B	47.90
GLAM-Swin-Unet B	**49.10**
GLAM-nogmsa-Swin-UperNet B	47.95
GLAM-Swin-UperNet B	**48.44**

Parameter and FLOPs overhead. The overhead due to the global tokens is controlled and proportional to the number of GLAM transformer blocks. This overhead brings higher performance gains than increasing the backbone size which validates the model architecture. Tab. 6 illustrates that the GLAM-Swin Base backbones show superior efficiency compared to their vanilla Large counterpart with a superior mIoU increase with respect to additional learnable parameters. The same analysis can be done with FLOPs overhead with a higher mIoU increase per extra-FLOP for GLAM-Swin Base compared to Swin Large.

Visualizations. Fig. Fig. 5 shows qualitative visualizations of the GLAM method. In Fig. Fig. 5a), we show GLAM attention maps for the highest resolution feature



a) Segmentation results and global attention of GLAM on ADE20K.



b) Segmentation results on Synapse.

Figure 5. Qualitative visualisations of GLAM. We show the ability of GLAM to model full contextual information in high-resolution feature maps on ADE20K (first row), and the ability of GLAM-nn-Former to accurately segment the stomach (in pink).

Table 6. Analysis of the relative mIoU increase with respect to extra learnable parameters and FLOPs compared to the standard Base and Large backbones.

backbone	#param.	↑ rel. mIoU / #param $\times 10^{-2}$	FLOPs	↑ rel. mIoU / FLOPs $\times 10^{-2}$
Swin-UperNet B	121	0	81G	0
Swin-UperNet L	234	0.4	180G	0.4
GLAM-Swin-UperNet B	197	**0.6**	99G	**2.5**

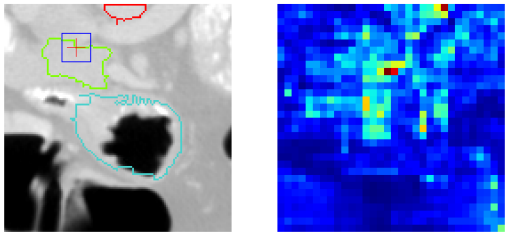


Figure 6. **Averaged GLAM attention map in 3D.** The information inside the blue window is ambiguous. To segment the voxel at the red cross, the model leverages long-range dependencies including neighbor organs. The pancreas is in green, the aorta in red, and the stomach in blue.

maps of a GLAM Swin-Unet model. Echoing observations in Fig. Fig. 1 in Cityscape, we can see that GLAM can model full-range interactions in this spatially-fine layer. This enables to exploit spatial relationships with other important structures (*e.g.* other sofas, arcades), which is not possible with the baseline Swin-Unet due to its limited window attention. We can notice the relevance of the GLAM segmentation. Furthermore, Fig. Fig. 6 shows the GLAM attention averaged over the axial direction for the red cross (pancreas). We can see that long-range dependencies are involved, with a much larger spatial extent than the local window (in blue), where attention is given to neighboring organs (stomach and aorta). The full context is crucial to properly segment complex organs with visual local ambiguities such as the pancreas. In Fig. Fig. 5b), we show segmentation results of GLAM-nn-Former for 3D medical image segmentation. We show the results on a given 2D slice. We can notice that GLAM nn-Former is qualitatively much better at segmenting the stomach (in pink) than nn-Former. This can be explained by the global interactions of our model, which enables it to better represent specific interactions between organs.

5. Conclusion

This paper introduces GLAM, a method for modeling full contextual interactions in multi-resolution transformer-based models. the GLAM transformer leverage learnable global tokens at each resolution level of the model, which allows a complete interaction of the tokens across the image regions, and is further equipped with a non-local upsampling module. Experiments show the large and consistent gain of GLAM when incorporated into several multi-resolution transformers (Swin-Unet, nn-Former, Swin) on diverse medical, street, or more general images. Future works includes applying the GLAM idea for modeling full contextual information on very high-resolution images or 3D medical volumes.

References

[1] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A non-local algorithm for image denoising. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pages 60–65, Washington, DC, USA, 2005. IEEE Computer Society.

[2] 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.

[3] Yao Chang, Hu Menghan, Zhai Guangtao, and Zhang Xiao-Ping. Transclaw u-net: Claw u-net with transformers for medical image segmentation, 2021.

[4] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. 2021.

[5] 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.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[7] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[8] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2020.

[9] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS 2021*, 2021.

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

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021.

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. 2019.

[15] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer, 2021.

[16] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4003–4012, 2020.

[17] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. 2019.

[18] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 2020.

[19] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. 2020.

[20] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2019.

[21] Bennett Landman, Zhoubing Xu, Igelsias Eugenio, Juan, Martin Styner, Thomas Robin, Langerak, and Arno Klein. Multi-atlas labeling beyond the cranial vault. *MICCAI*, 2015.

[22] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753, 2019.

[23] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers, 2021.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*, 2021.

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. pages 3431–3440.

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[27] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[28] Niki J. Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning (ICML)*, 2018.

[29] Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In *International Conference on Learning Representations*, 2020.

[30] Jiezhong Qiu, Hao Ma, Omer Levy, Scott Wen tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding, 2020.

[31] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2019.

[32] 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.

[33] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):640–651, 2017.

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers &; distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021.

[35] 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.

[36] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv e-prints*, pages arXiv–2006, 2020.

[37] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021.

[38] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE ICCV*, 2021.

[39] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

[40] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[41] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks, 2020.

[42] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, October 2021.

[43] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, and Jianfeng Gao. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. *ICCV 2021*, 2021.

[44] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating nested transformers. In *arXiv preprint arXiv:2105.12723*, 2021.

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[46] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021.

[47] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

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

[49] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020.